

1. [Preface](#)
2. [Additional Resources](#)
3. [Author Acknowledgements](#)
4. [Student Welcome Letter](#)
5. Sampling and Data
 1. [Sampling and Data](#)
 2. [Statistics](#)
 3. [Key Terms](#)
 4. [Data](#)
 5. [Sampling](#)
 6. [Variation](#)
 7. [Answers and Rounding Off](#)
 8. [Frequency](#)
 9. [Summary](#)
 10. [Practice: Sampling and Data](#)
6. Descriptive Statistics
 1. [Descriptive Statistics](#)
 2. [Displaying Data](#)
 3. [Stem and Leaf Graphs \(Stemplots\), Line Graphs and Bar Graphs](#)
 4. [Histograms](#)
 5. [Measures of the Location of the Data](#)
 6. [Measures of the Center of the Data](#)
 7. [Skewness and the Mean, Median, and Mode](#)
 8. [Understanding and Measuring Variability](#)
 9. [Measures of the Spread of the Data](#)
 10. [Summary of Formulas](#)
 11. [Normal Distribution: Introduction](#)
 12. [Variables](#)
7. The Normal Distribution
 1. [Introduction to Normal Distributions](#)

2. [Z-scores](#)
3. [The Standard Normal Distribution](#)
4. [The Normal Curve](#)
5. [Areas of Normal Distributions](#)
6. [Calculations of Probabilities](#)
7. [Normal Distribution: Calculations of Probabilities](#)
8. [z Scores with Critical Values](#)
8. The Central Limit Theorem
 1. [The Central Limit Theorem](#)
 2. [The Central Limit Theorem for Sample Means \(Averages\)](#)
 3. [The Central Limit Theorem for Sums](#)
 4. [Using the Central Limit Theorem](#)
 5. [Summary of Formulas](#)
 6. [Practice: The Central Limit Theorem](#)
 7. [Homework](#)
 8. [Review](#)
 9. [Lab 1: Central Limit Theorem \(Pocket Change\)](#)
 10. [Lab 2: Central Limit Theorem \(Cookie Recipes\)](#)
9. Confidence Intervals
 1. [Confidence Intervals](#)
 2. [Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal](#)
 3. [Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student-T](#)
 4. [Confidence Interval for a Population Proportion](#)
 5. [Summary of Formulas](#)
 6. [Practice 1: Confidence Intervals for Averages, Known Population Standard Deviation](#)
 7. [Practice 2: Confidence Intervals for Averages, Unknown Population Standard Deviation](#)
 8. [Practice 3: Confidence Intervals for Proportions](#)
 9. [Homework](#)

10. [Review](#)
 11. [Lab 1: Confidence Interval \(Home Costs\)](#)
 12. [Lab 2: Confidence Interval \(Place of Birth\)](#)
 13. [Lab 3: Confidence Interval \(Womens' Heights\)](#)
10. Linear Regression and Correlation
1. [Introduction to Bivariate Data](#)
 2. [Linear Regression and Correlation](#)
 3. [Linear Regression and Correlation: Linear Equations](#)
 4. [Linear Regression and Correlation: Slope and Y-Intercept of a Linear Equation](#)
 5. [Scatter Plots](#)
 6. [The Regression Equation](#)
 7. [Correlation Coefficient and Coefficient of Determination](#)
 8. [Testing the Significance of the Correlation Coefficient](#)
 9. [Prediction](#)
 10. [Outliers](#)
 11. [95% Critical Values of the Sample Correlation Coefficient Table](#)
 12. [Linear Regression and Correlation: Summary](#)
 13. [The Coefficient of Determination: Family Wealth and Student Achievement Scores](#)
11. Hypothesis Testing: Single Mean and Single Proportion
1. [Hypothesis Testing: Single Mean and Single Proportion](#)
 2. [Null and Alternate Hypotheses](#)
 3. [Outcomes and the Type I and Type II Errors](#)
 4. [Distribution Needed for Hypothesis Testing](#)
 5. [Assumption](#)
 6. [Rare Events](#)
 7. [Using the Sample to Support One of the Hypotheses](#)
 8. [Decision and Conclusion](#)
 9. [Additional Information](#)
 10. [Summary of the Hypothesis Test](#)

11. [Examples](#)
12. [Summary of Formulas](#)
13. [Effect Size](#)
12. Hypothesis Testing: Two Means, Paired Data, Two Proportions
 1. [Hypothesis Testing: Two Population Means and Two Population Proportions](#)
 2. [Comparing Two Independent Population Means with Unknown Population Standard Deviations](#)
 3. [Comparing Two Independent Population Means with Known Population Standard Deviations](#)
 4. [Comparing Two Independent Population Proportions](#)
 5. [Matched or Paired Samples](#)
 6. [Summary of Types of Hypothesis Tests](#)
 7. [Practice 1: Hypothesis Testing for Two Proportions](#)
 8. [Practice 2: Hypothesis Testing for Two Averages](#)
 9. [Review](#)
 10. [Lab: Hypothesis Testing for Two Means and Two Proportions](#)
13. The Chi-Square Distribution
 1. [The Chi-Square Distribution](#)
 2. [Notation](#)
 3. [Facts About the Chi-Square Distribution](#)
 4. [Goodness-of-Fit Test](#)
 5. [Test of Independence](#)
 6. [Summary of Formulas](#)
 7. [Practice 1: Goodness-of-Fit Test](#)
 8. [Practice 2: Contingency Tables](#)
 9. [Practice 3: Test of a Single Variance](#)
 10. [Review](#)
14. Appendix
 1. [Practice Final Exam 1](#)
 2. [Practice Final Exam 2](#)

3. [Data Sets](#)
4. Group Projects
 1. [Group Project: Univariate Data](#)
 2. [Group Project: Continuous Distributions and Central Limit Theorem](#)
 3. [Partner Project: Hypothesis Testing - Article](#)
 4. [Partner Project: Hypothesis Testing - Word Problem](#)
 5. [Group Project: Bivariate Data, Linear Regression, and Univariate Data](#)
5. Solution Sheets
 1. [Solution Sheet: Hypothesis Testing for Single Mean and Single Proportion](#)
 2. [Solution Sheet: Hypothesis Testing for Two Means, Paired Data, and Two Proportions](#)
 3. [Solution Sheet: The Chi-Square Distribution](#)
 4. [Solution Sheet: F Distribution and ANOVA](#)
6. [English Phrases Written Mathematically](#)
7. [Symbols and their Meanings](#)
8. [Formulas](#)
9. [Notes for the TI-83, 83+, 84 Calculator](#)
15. [Tables](#)

Preface

This module introduces the Connexions online textbook "Collaborative Statistics" by Barbara Illowsky and Susan Dean.

Welcome to *Collaborative Statistics*, presented by Connexions. The initial section below introduces you to Connexions. If you are familiar with Connexions, please skip to [About "Collaborative Statistics."](#)

About Connexions

Connexions Modular Content

Connexions (cnx.org) is an online, **open access** educational resource dedicated to providing high quality learning materials free online, free in printable PDF format, and at low cost in bound volumes through print-on-demand publishing. The *Collaborative Statistics* textbook is one of many **collections** available to Connexions users. Each **collection** is composed of a number of re-usable learning **modules** written in the Connexions XML markup language. Each module may also be re-used (or 're-purposed') as part of other collections and may be used outside of Connexions. Including *Collaborative Statistics*, Connexions currently offers over 6500 modules and more than 350 collections.

The modules of *Collaborative Statistics* are derived from the original paper version of the textbook under the same title, *Collaborative Statistics*. Each module represents a self-contained concept from the original work. Together, the modules comprise the original textbook.

Re-use and Customization

The [Creative Commons \(CC\) Attribution license](#) applies to all Connexions modules. Under this license, any module in Connexions may be used or modified for any purpose as long as proper attribution to the original author(s) is maintained. Connexions' authoring tools make re-use (or re-purposing) easy. Therefore, instructors anywhere are permitted to create customized versions of the *Collaborative Statistics* textbook by editing modules, deleting unneeded modules, and adding their own supplementary modules. Connexions' authoring tools keep track of these changes and maintain the CC license's required attribution to the original authors. This

process creates a new collection that can be viewed online, downloaded as a single PDF file, or ordered in any quantity by instructors and students as a low-cost printed textbook. To start building custom collections, please visit the help page, [“Create a Collection with Existing Modules”](#). For a guide to authoring modules, please look at the help page, [“Create a Module in Minutes”](#).

Read the book online, print the PDF, or buy a copy of the book.

To browse the *Collaborative Statistics* textbook online, visit the collection home page at cnx.org/content/col10522/latest. You will then have three options.

1. You may obtain a PDF of the entire textbook to print or view offline by clicking on the “Download PDF” link in the “Content Actions” box.
2. You may order a bound copy of the collection by clicking on the “Order Printed Copy” button.
3. You may view the collection modules online by clicking on the “Start >>” link, which takes you to the first module in the collection. You can then navigate through the subsequent modules by using their “Next >>” and “Previous >>” links to move forward and backward in the collection. You can jump to any module in the collection by clicking on that module’s title in the “Collection Contents” box on the left side of the window. If these contents are hidden, make them visible by clicking on “[show table of contents]”.

Accessibility and Section 508 Compliance

- For information on general Connexions accessibility features, please visit <http://cnx.org/content/m17212/latest/>.
- For information on accessibility features specific to the Collaborative Statistics textbook, please visit <http://cnx.org/content/m17211/latest/>.

Version Change History and Errata

- For a list of modifications, updates, and corrections, please visit <http://cnx.org/content/m17360/latest/>.

Adoption and Usage

- The Collaborative Statistics collection has been adopted and customized by a number of professors and educators for use in their classes. For a list of known versions and adopters, please visit <http://cnx.org/content/m18261/latest/>.

About “Collaborative Statistics”

Collaborative Statistics was written by Barbara Illowsky and Susan Dean, faculty members at De Anza College in Cupertino, California. The textbook was developed over several years and has been used in regular and honors-level classroom settings and in distance learning classes. Courses using this textbook have been articulated by the University of California for transfer of credit. The textbook contains full materials for course offerings, including expository text, examples, labs, homework, and projects. A Teacher’s Guide is currently available in print form and on the Connexions site at <http://cnx.org/content/col10547/latest/>, and supplemental course materials including additional problem sets and video lectures are available at <http://cnx.org/content/col10586/latest/>. The on-line text for each of these collections will meet the Section 508 standards for accessibility.

An on-line course based on the textbook was also developed by Illowsky and Dean. It has won an award as the best on-line California community college course. The on-line course will be available at a later date as a collection in Connexions, and each lesson in the on-line course will be linked to the on-line textbook chapter. The on-line course will include, in addition to expository text and examples, videos of course lectures in captioned and non-captioned format.

The original preface to the book as written by professors Illowsky and Dean, now follows:

This book is intended for introductory statistics courses being taken by students at two- and four-year colleges who are majoring in fields other than math or engineering. Intermediate algebra is the only prerequisite. The book focuses on applications of statistical knowledge rather than the theory

behind it. The text is named *Collaborative Statistics* because students learn best by **doing**. In fact, they learn best by working in small groups. The old saying “two heads are better than one” truly applies here.

Our emphasis in this text is on four main concepts:

- thinking statistically
- incorporating technology
- working collaboratively
- writing thoughtfully

These concepts are integral to our course. Students learn the best by actively participating, not by just watching and listening. Teaching should be highly interactive. Students need to be thoroughly engaged in the learning process in order to make sense of statistical concepts.

Collaborative Statistics provides techniques for students to write across the curriculum, to collaborate with their peers, to think statistically, and to incorporate technology.

This book takes students step by step. The text is interactive. Therefore, students can immediately apply what they read. Once students have completed the process of problem solving, they can tackle interesting and challenging problems relevant to today’s world. The problems require the students to apply their newly found skills. In addition, technology (TI-83 graphing calculators are highlighted) is incorporated throughout the text and the problems, as well as in the special group activities and projects. The book also contains labs that use real data and practices that lead students step by step through the problem solving process.

At De Anza, along with hundreds of other colleges across the country, the college audience involves a large number of ESL students as well as students from many disciplines. The ESL students, as well as the non-ESL students, have been especially appreciative of this text. They find it extremely readable and understandable. *Collaborative Statistics* has been used in classes that range from 20 to 120 students, and in regular, honor, and distance learning classes.

Susan Dean

Barbara Illowsky

Additional Resources

This module catalogs several of the resources available for teachers and students using the Collaborative Statistics (col10522) textbook and its derivatives. This module provides links to the complementary teacher's guide, supplemental materials including video lectures and additional problem sets, accessibility information, collection version history and errata, and a list of related works and teachers who have adopted them for their courses.

Additional Resources Currently Available

- [Glossary](#)
- [View or Download This Textbook Online](#)
- [Collaborative Statistics Teacher's Guide](#)
- [Supplemental Materials](#)
- [Video Lectures](#)
- [Version History](#)
- [Textbook Adoption and Usage](#)
- [Additional Technologies and Notes](#)
- [Accessibility and Section 508 Compliance](#)

The following section describes some additional resources for learners and educators. These modules and collections are all available on the Connexions website (<http://cnx.org/>) and can be viewed online, downloaded, printed, or ordered as appropriate.

Glossary

This module contains the entire glossary for the Collaborative Statistics textbook collection (col10522) since its initial release on 15 July 2008. The glossary is located at <http://cnx.org/content/m16129/latest/>.

Below are links to additional resources:

Link to the Statistics Glossary by Dr. Philip Stark, UC Berkeley

[http:// statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm](http://statistics.berkeley.edu/~stark/SticiGui/Text/gloss.htm)

Link to Wikipedia

http:// <http://www.wikipedia.org/>

(Search on "Glossary of probability and statistics.")

View or Download This Textbook Online

The complete contents of this book are available at no cost on the Connexions website at <http://cnx.org/content/col10522/latest/>. Anybody can view this content free of charge either as an online e-book or a downloadable PDF file. A low-cost printed version of this textbook is also available [here](#).

Collaborative Statistics Teacher's Guide

A complementary Teacher's Guide for Collaborative statistics is available through Connexions at <http://cnx.org/content/col10547/latest/>. The Teacher's Guide includes suggestions for presenting concepts found throughout the book as well as recommended homework assignments. A low-cost printed version of this textbook is also available [here](#).

Supplemental Materials

This companion to Collaborative Statistics provides a number of additional resources for use by students and instructors based on the award winning [Elementary Statistics Sofia online course](#), also by textbook authors Barbara Illowsky and Susan Dean. This content is designed to complement the textbook by providing video tutorials, course management materials, and sample problem sets. The Supplemental Materials collection can be found at <http://cnx.org/content/col10586/latest/>.

Video Lectures

- [Video Lecture 1: Sampling and Data](#)
- [Video Lecture 2: Descriptive Statistics](#)
- [Video Lecture 3: Probability Topics](#)
- [Video Lecture 4: Discrete Distributions](#)
- [Video Lecture 5: Continuous Random Variables](#)
- [Video Lecture 6: The Normal Distribution](#)
- [Video Lecture 7: The Central Limit Theorem](#)
- [Video Lecture 8: Confidence Intervals](#)
- [Video Lecture 9: Hypothesis Testing with a Single Mean](#)
- [Video Lecture 10: Hypothesis Testing with Two Means](#)
- [Video Lecture 11: The Chi-Square Distribution](#)

- [Video Lecture 12: Linear Regression and Correlation](#)

Version History

This module contains a listing of changes, updates, and corrections made to the Collaborative Statistics textbook collection (col10522) since its initial release on 15 July 2008. The Version History is located at <http://cnx.org/content/m17360/latest/>.

Textbook Adoption and Usage

This module is designed to track the various derivations of the Collaborative Statistics textbook and its various companion resources, as well as keep track of educators who have adopted various versions for their courses. New adopters are encouraged to provide their contact information and describe how they will use this book for their courses. The goal is to provide a list that will allow educators using this book to collaborate, share ideas, and make suggestions for future development of this text. The Adoption and Usage module is located at <http://cnx.org/content/m18261/latest/>.

Additional Technologies

In order to provide the most flexible learning resources possible, we invite collaboration from all instructors wishing to create customized versions of this content for use with other technologies. For instance, you may be interested in creating a set of instructions similar to this collection's calculator notes. If you would like to contribute to this collection, please use the contact the authors with any ideas or materials you have created.

Accessibility and Section 508 Compliance

- For information on general Connexions accessibility features, please visit <http://cnx.org/content/m17212/latest/>.
- For information on accessibility features specific to the Collaborative Statistics textbook, please visit <http://cnx.org/content/m17211/latest/>.

Author Acknowledgements

This module contains the author acknowledgements for the Collaborative Statistics textbook/collection.

For this second edition, we appreciate the tremendous feedback from De Anza College colleagues and students, as well as from the dozens of faculty around the world who taught out of the first and preliminary editions. We have updated Collaborative Statistics with contributions from many faculty and students. We especially thank Roberta Bloom, who wrote new problems and additional text.

So many students and colleagues have contributed to the text, both the hard copy and open version. We thank the following people for their contributions to the first and/or second editions.

At De Anza College:

Dr. Inna Grushko (deceased), who wrote the glossary; Diane Mathios, who checked every homework problem in the first edition; Kathy Plum, Lenore Desilets, Charles Klein, Janice Hector, Frank Snow, Dr. Lisa Markus, Dr. Vladimir Logvinenko (deceased), Mo Geraghty, Rupinder Sekhon, Javier Rueda, Carol Olmstead; Also, Dr. Jim Lucas and Valerie Hauber of De Anza's Office of Institutional Research, Mary Jo Kane of Health Services; and the thousands of students who have used this text. Many of the students gave us permission to include their outstanding word problems as homework.

Additional thanks:

Dr. Larry Green of Lake Tahoe Community College, Terrie Teegarden of San Diego Mesa College, Ann Flanigan of Kapiolani Community College, Birgit Aquilonius of West Valley College.

The conversion from a for-profit hard copy text to a free open textbook is the result of many individuals and organizations. We particularly thank Dr. Martha Kanter, Hal Plotkin, Dr. Judy Baker, Dr. Robert Maxfield of Maxfield Foundation, Hewlett Foundation, and Connexions.

Finally, we owe much to Frank, Jeffrey, and Jessica Dean and to Dan, Rachel, Matthew, and Rebecca Illowsky, who encouraged us to continue

with our work and who had to hear more than their share of “I’m sorry, I can’t” and “Just a minute, I’m working.”

Student Welcome Letter

Dear Student:

Have you heard others say, “You’re taking statistics? That’s the hardest course I ever took!” They say that, because they probably spent the entire course confused and struggling. They were probably lectured to and never had the chance to experience the subject. You will not have that problem. Let’s find out why.

There is a Chinese Proverb that describes our feelings about the field of statistics:

I HEAR, AND I FORGET

I SEE, AND I REMEMBER

I DO, AND I UNDERSTAND

Statistics is a “do” field. In order to learn it, you must “do” it. We have structured this book so that you will have hands-on experiences. They will enable you to truly understand the concepts instead of merely going through the requirements for the course.

What makes this book different from other texts? First, we have eliminated the drudgery of tedious calculations. You might be using computers or graphing calculators so that you do not need to struggle with algebraic manipulations. Second, this course is taught as a collaborative activity. With others in your class, you will work toward the common goal of learning this material.

Here are some hints for success in your class:

- Work hard and work every night.
- Form a study group and learn together.
- Don’t get discouraged - you can do it!
- As you solve problems, ask yourself, “Does this answer make sense?”
- Many statistics words have the same meaning as in everyday English.

- Go to your teacher for help as soon as you need it.
- Don't get behind.
- Read the newspaper and ask yourself, "Does this article make sense?"
- Draw pictures - they truly help!

Good luck and don't give up!

Sincerely,
Susan Dean and Barbara Illowsky

De Anza College
21250 Stevens Creek Blvd.
Cupertino, California 95014

Sampling and Data

This module provides a brief introduction to the field of statistics, including examples of how these topics shows up in a variety of real-life examples.

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.
- Create and interpret frequency tables.

Introduction

You are probably asking yourself the question, "When and where will I use statistics?". If you read any newspaper or watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a news program on television, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques to analyze the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data are.

Statistics

This module introduces the concept of statistics, specifically the ability to use statistics to describe data (descriptive statistics) as well as draw conclusions (inferential statistics). An optional classroom exercise is included.

The science of [statistics](#) deals with the collection, analysis, interpretation, and presentation of [data](#). We see and use data in our everyday lives.

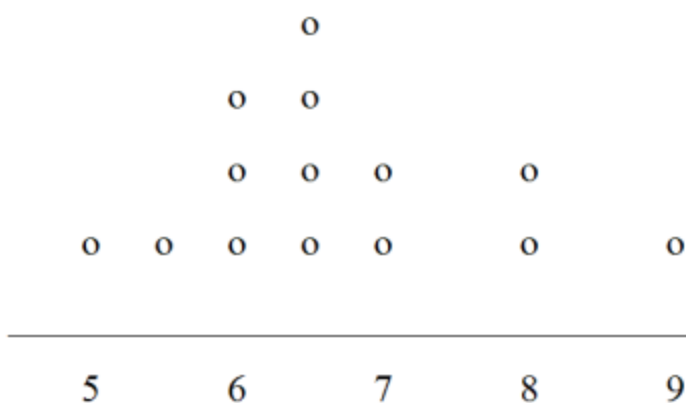
Optional Collaborative Classroom Exercise

In your classroom, try this exercise. Have class members write down the average time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, consider the following data:

5 5.5 6 6 6 6.5 6.5 6.5 6.5 7 7 8 8 9

The dot plot for this data would be as follows:

Frequency of Average Time (in Hours) Spent Sleeping per Night



Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?

Where do your data appear to cluster? How could you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that the conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Levels of Measurement and Statistical Operations

The way a set of data is measured is called its level of measurement. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are (from lowest to highest level):

- Nominal scale level
- Ordinal scale level
- Interval scale level
- Ratio scale level

Data that is measured using a **nominal scale** is qualitative. Categories, colors, names, labels and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful.

Smartphone companies are another example of nominal scale data. Some examples are Sony, Motorola, Nokia, Samsung and Apple. This is just a list and there is no agreed upon order. Some people may favor Apple but that is a matter of opinion. Nominal scale data cannot be used in calculations.

Data that is measured using an **ordinal scale** is similar to nominal scale data but there is a big difference. The ordinal scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure differences between the data.

Another example using the ordinal scale is a cruise survey where the responses to questions about the cruise are “excellent,” “good,” “satisfactory” and “unsatisfactory.” These responses are ordered from the most desired response by the cruise lines to the least desired. But the differences between two pieces of data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the **interval scale** is similar to ordinal level data because it has a definite ordering but there is a difference between data. The differences between interval scale data can be measured though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40 degrees is equal to 100 degrees minus 60 degrees. Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -10° F and -15° C exist and are colder than 0.

Interval level data can be used in calculations but one type of comparison cannot be done. Eighty degrees C is not 4 times as hot as 20° C (nor is 80° F 4 times as hot as 20° F). There is no meaning to the ratio of 80 to 20 (or 4 to 1).

Data that is measured using the **ratio scale** takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data but, in addition, it has a 0 point and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The exams were machine-graded.

The data can be put in order from lowest to highest: 20, 68, 80, 92.

The differences between the data have meaning. The score 92 is more than the score 68 by 24 points.

Ratios can be calculated. The smallest score for ratio data is 0. So 80 is 4 times 20. The score of 80 is 4 times better than the score of 20.

Exercises

What type of measure scale is being used? Nominal, Ordinal, Interval or Ratio.

1. High school men soccer players classified by their athletic ability:
Superior, Average, Above average.
2. Baking temperatures for various main dishes: 350, 400, 325, 250, 300
3. The colors of crayons in a 24-crayon box.
4. Social security numbers.
5. Incomes measured in dollars
6. A satisfaction survey of a social website by number: 1 = very satisfied, 2 = somewhat satisfied, 3 = not satisfied.
7. Political outlook: extreme left, left-of-center, right-of-center, extreme right.
8. Time of day on an analog watch.
9. The distance in miles to the closest grocery store.
10. The dates 1066, 1492, 1644, 1947, 1944.

11. The heights of 21 – 65 year-old women.
12. Common letter grades A, B, C, D, F.

Answers 1. ordinal, 2. interval, 3. nominal, 4. nominal, 5. ratio, 6. ordinal, 7. nominal, 8. interval, 9. ratio, 10. interval, 11. ratio, 12. ordinal

Glossary

Data

A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Statistic

A numerical characteristic of the sample. A statistic estimates the corresponding population parameter. For example, the average number of full-time students in a 7:30 a.m. class for this term (statistic) is an estimate for the average number of full-time students in any class this term (parameter).

Key Terms

This module introduces a number of key terms related to statistical sampling and data.

In statistics, we generally want to study a **population**. You can think of a population as an entire collection of persons, things, or objects under study. To study the larger population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000 to 2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that is a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a number that is a property of the population. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, notated by capital letters like X and Y , is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. Datum is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtained scores of 86, 75, and 92, you calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

Note:

Mean and Average

The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean" and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example:

Exercise:

Problem:

Define the key terms from the following study: We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly survey 100 first year students at the college. Three of those students spent \$150, \$200, and \$225, respectively.

Solution:

The **population** is all first year students attending ABC College this term.

The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).

The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term.

The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample.

The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** are the dollar amounts spent by the first year students. Examples of the data are \$150, \$200, and \$225.

Optional Collaborative Classroom Exercise

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

Glossary

Average

A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Data

A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Proportion

- As a number: A proportion is the number of successes divided by the total number in the sample.
- As a probability distribution: Given a binomial random variable (RV), $X \sim B(n, p)$, consider the ratio of the number X of successes in n Bernoulli trials to the number n of trials. $P = \frac{X}{n}$. This new RV is called a proportion, and if the number of trials, n , is large enough, $P \sim N\left(p, \frac{pq}{n}\right)$.

Data

This module introduces the concepts of qualitative data, quantitative continuous data, and quantitative discrete data as used in statistics. Sample problems are included.

Data may come from a population or from a sample. Small letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- Qualitative
- Quantitative

Qualitative data are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and the number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get 0, 1, 2, 3, etc.

All data that are the result of measuring are **quantitative continuous data** assuming that we can measure accurately. Measuring angles in radians might result in the numbers $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, π , $\frac{3\pi}{4}$, etc. If you and your friends carry backpacks with books in them to school, the numbers of books in the

backpacks are discrete data and the weights of the backpacks are continuous data.

Note: In this course, the data used is mainly quantitative. It is easy to calculate statistics (like the mean or proportion) from numbers. In the chapter **Descriptive Statistics**, you will be introduced to stem plots, histograms and box plots all of which display quantitative data. Qualitative data is discussed at the end of this section through graphs.

Example:

Data Sample of Quantitative Discrete Data

The data are the number of books students carry in their backpacks. You sample five students. Two students carry 3 books, one student carries 4 books, one student carries 2 books, and one student carries 1 book. The numbers of books (3, 4, 2, and 1) are the quantitative discrete data.

Example:

Data Sample of Quantitative Continuous Data

The data are the weights of the backpacks with the books in it. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are quantitative continuous data because weights are measured.

Example:

Data Sample of Qualitative Data

The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative data.

Note: You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

Example:

Exercise:

Problem:

Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of."

1. The number of pairs of shoes you own.
2. The type of car you drive.
3. Where you go on vacation.
4. The distance it is from your home to the nearest grocery store.
5. The number of classes you take per school year.
6. The tuition for your classes
7. The type of calculator you use.
8. Movie ratings.
9. Political party preferences.
10. Weight of sumo wrestlers.
11. Amount of money won playing poker.
12. Number of correct answers on a quiz.
13. Peoples' attitudes toward the government.
14. IQ scores. (This may cause some discussion.)

Solution:

Items 1, 5, 11, and 12 are quantitative discrete; items 4, 6, 10, and 14 are quantitative continuous; and items 2, 3, 7, 8, 9, and 13 are qualitative.

Qualitative Data Discussion

Below are tables of part-time vs full-time students at De Anza College in Cupertino, CA and Foothill College in Los Altos, CA for the Spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

	Number	Percent
Full-time	9,200	40.9%
Part-time	13,296	59.1%
Total	22,496	100%

De Anza College

	Number	Percent
Full-time	4,059	28.6%
Part-time	10,124	71.4%

Total	14,183	100%
-------	--------	------

Foothill College

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning what graphs to use. Below are pie charts and bar graphs, two graphs that are used to display qualitative data.

In a **pie chart**, categories of data are represented by wedges in the circle and are proportional in size to the percent of individuals in each category.

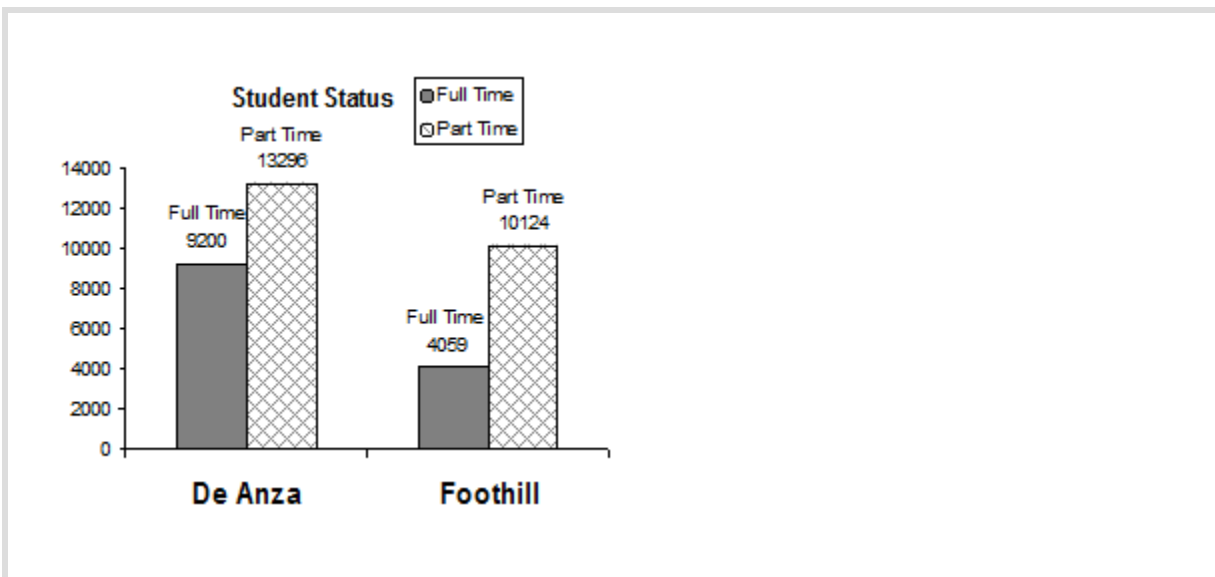
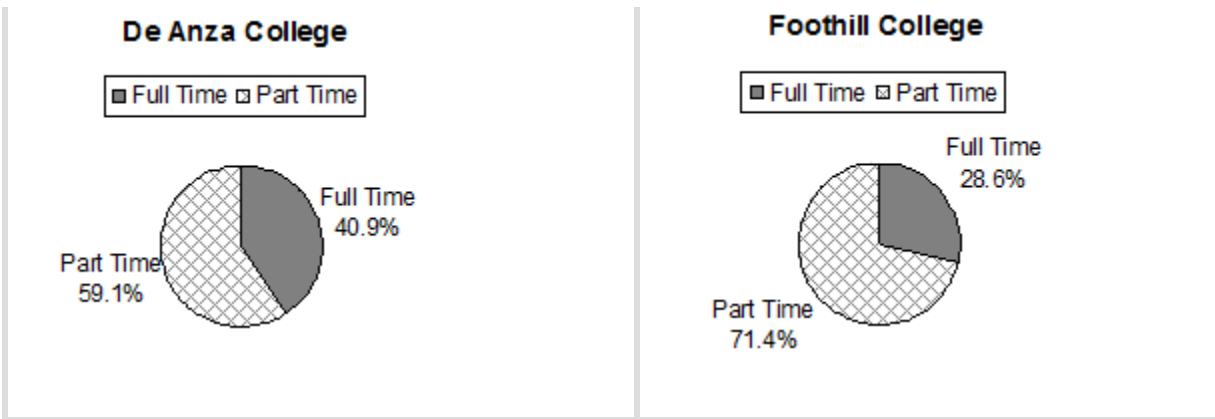
In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.

A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at the graphs and determine which graph (pie or bar) you think displays the comparisons better. This is a matter of preference.

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the "best" graph depending on the data and the context. Our choice also depends on what we are using the data for.

--	--

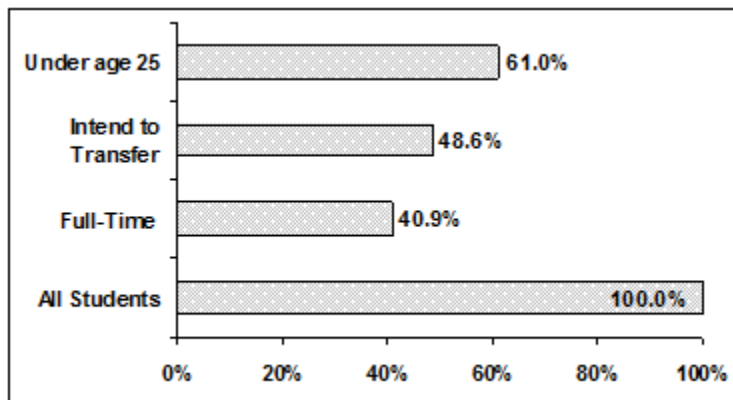


Percentages That Add to More (or Less) Than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Characteristic/Category	Percent
Full-time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

De Anza College Spring 2010

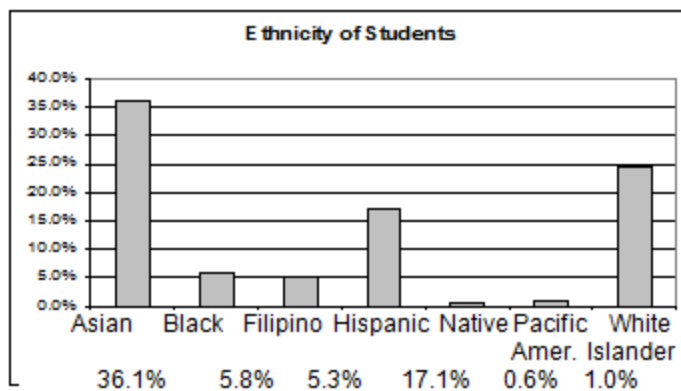


Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. Create a bar graph and not a pie chart.

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

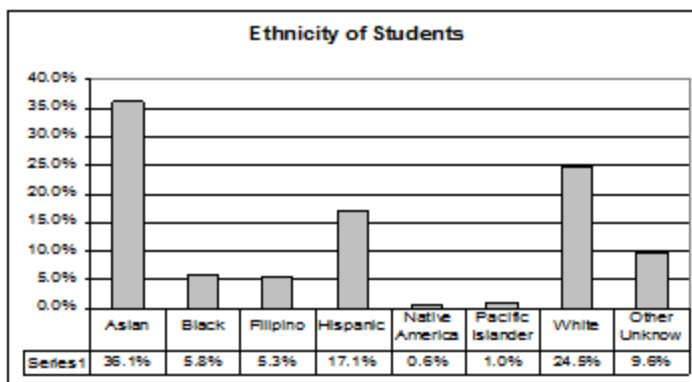
Missing Data: Ethnicity of Students De Anza College Fall Term 2007
(Census Day)



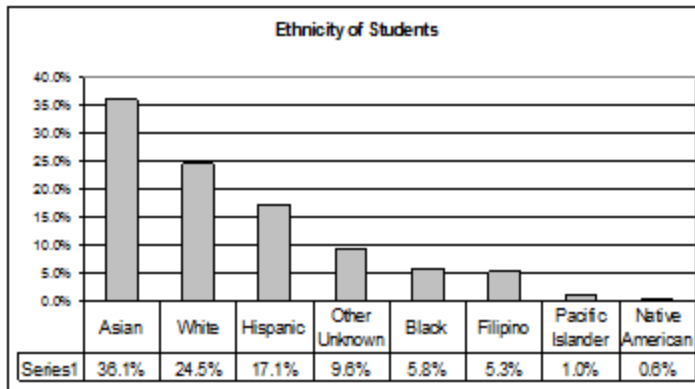
Bar graph Without Other/Unknown Category

The following graph is the same as the previous graph but the "Other/Unknown" percent (9.6%) has been added back in. The "Other/Unknown" category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0% particularly). This is important to know when we think about what the data are telling us.

This particular bar graph can be hard to understand visually. The graph below it is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.



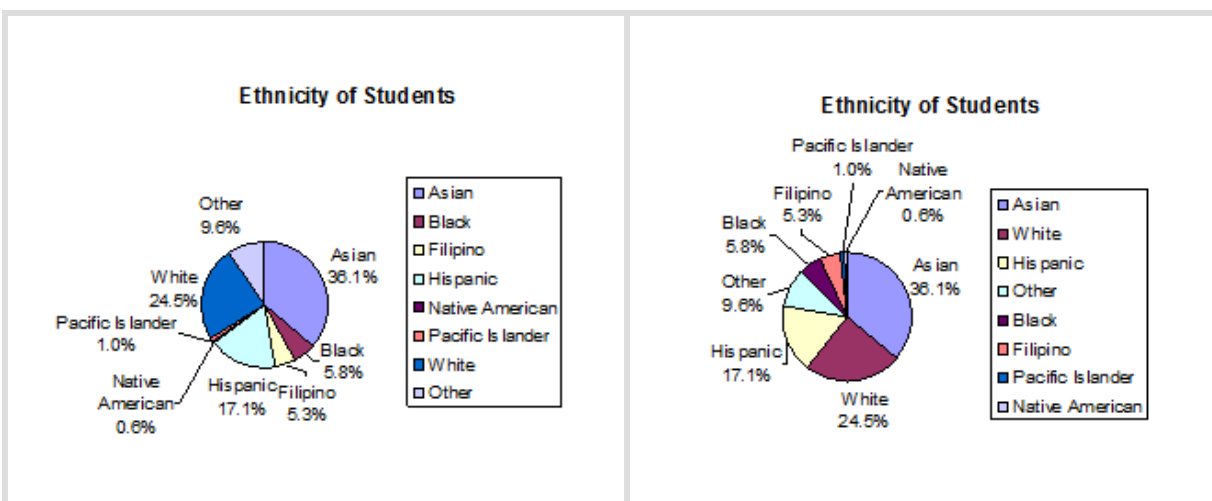
Bar Graph With Other/Unknown Category



Pareto Chart With Bars Sorted By Size

Pie Charts: No Missing Data

The following pie charts have the "Other/Unknown" category added back in (since the percentages must add to 100%). The chart on the right is organized having the wedges by size and makes for a more visually informative graph than the unsorted, alphabetical graph on the left.



Glossary

Continuous Random Variable

A random variable (RV) whose outcomes are measured.

Example:

The height of trees in the forest is a continuous RV.

Data

A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Discrete Random Variable

A random variable (RV) whose outcomes are counted.

Qualitative Data

See [Data](#).

Quantitative Data

See [Data](#).

Sampling

This module introduces the concept of statistical sampling. Students are taught the difference between a simple random sample, stratified sample, cluster sample, systematic sample, and convenience sample. Example problems are provided, including an optional classroom activity.

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods.

There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of n individuals is equally likely to be chosen by any other group of n individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size 3 from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out 3 names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number as shown below.

ID	Name
00	Anselmo

ID	Name
01	Bautista
02	Bayani
03	Cheng
04	Cuarismo
05	Cunningham
06	Fontecha
07	Hong
08	Hoobler
09	Jiao
10	Khan
11	King
12	Legeny
13	Lundquist
14	Macierz
15	Motogawa
16	Okimoto
17	Patel

ID	Name
18	Price
19	Quizon
20	Reyes
21	Roquero
22	Roth
23	Rowell
24	Salangsang
25	Slade
26	Stracher
27	Tallai
28	Tran
29	Wai
30	Wood

Class Roster

Lisa can either use a table of random numbers (found in many statistics books as well as mathematical handbooks) or a calculator or computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are:

.94360 .99832 .14669 .51470 .40581 .73381 .04399

Lisa reads two-digit groups until she has chosen three class members (that is, she reads .94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers .94360 and .99832 do not contain appropriate two digit numbers. However the third random number, .14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cunningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, and Cunningham, and Cuarismo.

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. For example, divide your college faculty by department. The departments are the clusters. Number each department and then choose

four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n th piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1 - 20,000 and then use a simple random sample to pick a number that represents the first name of the sample. Then choose every 50th name thereafter until you have a total of 400 names (you might have to go back to the of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is nonrandom is convenience sampling.

Convenience sampling involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favors certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (for example, they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once using with replacement is very low.

For example, in a college population of 10,000 people, suppose you want to randomly pick a sample of 1000 for a survey. **For any particular sample of 1000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions $999/10,000$ and $999/9,999$. For accuracy, carry the decimal answers to 4 place decimals. To 4 decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement only becomes a mathematics issue when the population is small which is not that common. For example, if the population is 25 people, the sample is 10 and you are sampling **with replacement for any particular sample**,

- the chance of picking the first person is 10 out of 25 and a different second person is 9 out of 25 (you replace the first person).

If you sample **without replacement**,

- the chance of picking the first person is 10 out of 25 and then the second person (which is different) is 9 out of 24 (you do not replace the first person).

Compare the fractions $9/25$ and $9/24$. To 4 decimal places, $9/25 = 0.3600$ and $9/24 = 0.3750$. To 4 decimal places, these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, a **sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Example:

Exercise:

Problem:

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

1. A soccer coach selects 6 players from a group of boys aged 8 to 10, 7 players from a group of boys aged 11 to 12, and 3 players from a group of boys aged 13 to 14 to form a recreational soccer team.
2. A pollster interviews all human resource personnel in five different high tech companies.
3. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
4. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.

5. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
6. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Solution:

1. stratified
2. cluster
3. stratified
4. systematic
5. simple random
6. convenience

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will seem natural.

Example:

Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task.

Suppose we take two different samples.

First, we use convenience sampling and survey 10 students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class. The amount of money they spend is as follows:

\$128 \$87 \$173 \$116 \$130 \$204 \$147 \$189 \$93 \$153

The second sample is taken by using a list from the P.E. department of senior citizens who take P.E. classes and taking every 5th senior citizen on the list, for a total of 10 senior citizens. They spend:

\$50 \$40 \$36 \$15 \$50 \$100 \$40 \$53 \$22 \$22

Exercise:

Problem:

Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Solution:

No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average part-time student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

Exercise:

Problem:

Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Solution:

No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of

part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he/she has a corresponding number. The students spend:
\$180 \$50 \$150 \$85 \$260 \$75 \$180 \$200 \$200 \$150

Exercise:

Problem: Is the sample biased?

Solution:

The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Optional Collaborative Classroom Exercise

Exercise:

Problem:

As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

1. To find the average GPA of all students in a university, use all honor students at the university as the sample.
2. To find out the most popular cereal among young people under the age of 10, stand outside a large supermarket for three hours and speak to every 20th child under age 10 who enters the supermarket.

3. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
4. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
5. To determine the average cost of a two day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

Variation

This module discusses statistical variability within data and samples. Students will be given the opportunity to see this variability in action through participation in an optional classroom exercise. This module also has a section that discusses Critical Evaluation.

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8 16.1 15.2 14.8 15.8 15.9 16.0 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

Variation in Samples

It was mentioned previously that two or more [samples](#) from the same [population](#), taken randomly, and having close to the same characteristics of the population are different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the

same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This **variability in samples** cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1200 to 1500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariably biased because people choose to respond or not.

Optional Collaborative Classroom Exercise

Exercise:

Problem:

Divide into groups of two, three, or four. Your instructor will give each group one 6-sided die. **Try this experiment twice.** Roll one fair die (6-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get below ("frequency" is the number of times a particular face of the die occurs):

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

First Experiment (20 rolls)

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

Second Experiment (20 rolls)

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? (Answer yes or no.) Why or why not?

Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

Critical Evaluation

We need to critically evaluate the statistical studies we read about and analyze before accepting the results of the study. Common problems to be aware of include

- **Problems with Samples:** A sample should be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- **Self-Selected Samples:** Responses only by people who choose to respond, such as call-in surveys are often unreliable.
- **Sample Size Issues:** Samples that are too small may be unreliable. Larger samples are better if possible. In some situations, small samples are unavoidable and can still be used to draw conclusions, even though larger samples are better. Examples: Crash testing cars, medical testing for rare conditions.
- **Undue influence:** Collecting data or asking questions in a way that influences the response.
- **Non-response or refusal of subject to participate:** The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- **Causality:** A relationship between two variables does not mean that one causes the other to occur. They may both be related (correlated) because of their relationship through a different variable.
- **Self-Funded or Self-Interest Studies:** A study performed by a person or organization in order to support their claim. Is the study impartial?

Read the study carefully to evaluate the work. Do not automatically assume that the study is good but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.

- **Misleading Use of Data:** Improperly displayed graphs, incomplete data, lack of context.
- **Confounding:** When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

Glossary

Population

The collection, or set, of all individuals, objects, or measurements whose properties are being studied.

Sample

A portion of the population under study. A sample is representative if it characterizes the population being studied.

Answers and Rounding Off

This module briefly explains the correct way to round off answers when working with statistical data.

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round only the final answer. Do not round any intermediate results, if possible. If it becomes necessary to round intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores 4, 6, 9 is 6.3, rounded to the nearest tenth, because the data are whole numbers. Most answers will be rounded in this manner.

It is not necessary to reduce most fractions in this course. Especially in [Probability Topics](#), the chapter on probability, it is more helpful to leave an answer as an unreduced fraction.

Frequency

This module introduces the concepts of frequency, relative frequency, and cumulative relative frequency, and the relationship between these measures. Students will have the opportunity to interpret data through the sample problems provided.

Frequency Distributions

To make use of data we need to summarize them into a form that is easy to talk about. The first step is to obtain a frequency, a list of how often a score appears.

Twenty students were asked how many hours they worked per day. Their responses, in hours, are listed below:

5 6 3 3 2 4 7 5 2 3 5 6 5 4 4 3 5 2 5 3

Below is a frequency table listing the different data values in ascending order and their frequencies.

DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6
6	2
7	1

Frequency Table of Student Work Hours

A **frequency** is the number of times a given datum occurs in a data set. According to the table above, there are three students who work 2 hours, five

students who work 3 hours, etc. The total of the frequency column, 20, represents the total number of students included in the sample.

Frequency Distribution

A Frequency Distribution is a tool for organizing data. It lists all the values a variable can take and how often each occurs. A frequency distribution is often just referred to as a distribution.

A **relative frequency** is the fraction or proportion of times an answer occurs. To find the relative frequencies, divide each frequency by the total number of students in the sample - in this case, 20. Relative frequencies can be written as fractions, percents, or decimals. The formula for a **percent** or **proportion** is the number of hits (whatever you are looking for) divided by the total. Looking at table 2, what percent of the scores were exactly five? The number of five's is 6 and there are a total of 20 scores so 6 divided by 20 gives a proportion of .30. Multiply that by 100 and add a percent sign and the answer is 30%.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

Frequency Table of Student Work Hours w/ Relative Frequency

The sum of the relative frequency column is $\frac{20}{20}$, or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	$0.15 + 0.25 = 0.40$
4	3	$\frac{3}{20}$ or 0.15	$0.40 + 0.15 = 0.55$
5	6	$\frac{6}{20}$ or 0.30	$0.55 + 0.30 = 0.85$
6	2	$\frac{2}{20}$ or 0.10	$0.85 + 0.10 = 0.95$
7	1	$\frac{1}{20}$ or 0.05	$0.95 + 0.05 = 1.00$

Frequency Table of Student Work Hours w/ Relative and Cumulative Relative Frequency

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

Note: Because of rounding, the relative frequency column may not always sum to one and the last entry in the cumulative relative frequency column may not be

one. However, they each should be close to one.

The following table represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95 - 61.95	5	$\frac{5}{100} = 0.05$	0.05
61.95 - 63.95	3	$\frac{3}{100} = 0.03$	$0.05 + 0.03 =$ 0.08
63.95 - 65.95	15	$\frac{15}{100} = 0.15$	$0.08 + 0.15 =$ 0.23
65.95 - 67.95	40	$\frac{40}{100} = 0.40$	$0.23 + 0.40 =$ 0.63
67.95 - 69.95	17	$\frac{17}{100} = 0.17$	$0.63 + 0.17 =$ 0.80
69.95 - 71.95	12	$\frac{12}{100} = 0.12$	$0.80 + 0.12 =$ 0.92
71.95 - 73.95	7	$\frac{7}{100} = 0.07$	$0.92 + 0.07 =$ 0.99

	Total = 100	Total = 1.00	
--	-------------	--------------	--

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
73.95 - 75.95	1	$\frac{1}{100} = 0.01$	$0.99 + 0.01 =$ 1.00
	Total = 100	Total = 1.00	

Frequency Table of Soccer Player Height

The data in this table has been **grouped** into the following intervals:

- 59.95 - 61.95 inches
- 61.95 - 63.95 inches
- 63.95 - 65.95 inches
- 65.95 - 67.95 inches
- 67.95 - 69.95 inches
- 69.95 - 71.95 inches
- 71.95 - 73.95 inches
- 73.95 - 75.95 inches

Note: This example is used again in the [Descriptive Statistics](#) chapter, where the method used to compute the intervals will be explained.

In this sample, there are **5** players whose heights are between 59.95 - 61.95 inches, **3** players whose heights fall within the interval 61.95 - 63.95 inches, **15** players whose heights fall within the interval 63.95 - 65.95 inches, **40** players whose heights fall within the interval 65.95 - 67.95 inches, **17** players whose heights fall within the interval 67.95 - 69.95 inches, **12** players whose heights fall within the interval 69.95 - 71.95, 7 players whose height falls within the interval 71.95 - 73.95, and **1** player whose height falls within the interval 73.95 - 75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Example:

Exercise:

Problem:

From the table, find the percentage of heights that are less than 65.95 inches.

Solution:

If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are $5 + 3 + 15 = 23$ males whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then $\frac{23}{100}$ or 23%. This percentage is the cumulative relative frequency entry in the third row.

Example:

Exercise:

Problem:

From the table, find the percentage of heights that fall between 61.95 and 65.95 inches.

Solution:

Add the relative frequencies in the second and third rows: $0.03 + 0.15 = 0.18$ or 18%.

Example:

Exercise:

Problem:

Use the table of heights of the 100 male semiprofessional soccer players. Fill in the blanks and check your answers.

1. The percentage of heights that are from 67.95 to 71.95 inches is:
2. The percentage of heights that are from 67.95 to 73.95 inches is:
3. The percentage of heights that are more than 65.95 inches is:
4. The number of players in the sample who are between 61.95 and 71.95 inches tall is:
5. What kind of data are the heights?
6. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

Solution:

1. 29%
2. 36%
3. 77%
4. 87
5. quantitative continuous
6. get rosters from each team and choose a simple random sample from each

Glossary

Frequency

The number of times a value of the data occurs.

Relative Frequency

The ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes.

Cumulative Relative Frequency

The term applies to an ordered set of observations from smallest to largest. The Cumulative Relative Frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Summary

This module provides an outline/review of key concepts related to statistical sampling and data.

Statistics

- Deals with the collection, analysis, interpretation, and presentation of data

Probability

- Mathematical tool used to study randomness

Key Terms

- Population
- Parameter
- Sample
- Statistic
- Variable
- Data

Types of Data

- Quantitative Data (a number)
 - Discrete (You count it.)
 - Continuous (You measure it.)
- Qualitative Data (a category, words)

Sampling

- **With Replacement:** A member of the population may be chosen more than once
- **Without Replacement:** A member of the population may be chosen only once

Random Sampling

- Each member of the population has an equal chance of being selected

Sampling Methods

- Random
 - Simple random sample
 - Stratified sample
 - Cluster sample
 - Systematic sample
- Not Random
 - Convenience sample

Frequency (freq. or f)

- The number of times an answer occurs

Relative Frequency (rel. freq. or RF)

- The proportion of times an answer occurs
- Can be interpreted as a fraction, decimal, or percent

Cumulative Relative Frequencies (cum. rel. freq. or cum RF)

- An accumulation of the previous relative frequencies

Practice: Sampling and Data

This module provides an opportunity for students to practice concepts related to statistical sampling and data. Given a sample data set, the student will practice constructing frequency tables, differentiating between key terms, and comparing sampling techniques.

Student Learning Outcomes

- The student will construct frequency tables.
- The student will differentiate between key terms.
- The student will compare sampling techniques.

Given

Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average(mean) length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A 3 4 11 15 16 17 22 44 37 16 14 24 25 15 26 27 33 29 35 44
13 21 22 10 12 8 40 32 26 27 31 34 29 17 8 24 18 47 33 34

Researcher B 3 14 11 5 16 17 28 41 31 18 14 14 26 25 21 22 31 2 35 44 23
21 21 16 12 18 41 22 16 25 33 34 29 13 18 24 23 42 33 29

Organize the Data

Complete the tables below using the data provided.

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5 - 6.5			
6.5 - 12.5			
12.5 - 18.5			
18.5 - 24.5			
24.5 - 30.5			
30.5 - 36.5			
36.5 - 42.5			
42.5 - 48.5			

Researcher A

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5 - 6.5			
6.5 - 12.5			
12.5 - 18.5			

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
18.5 - 24.5			
24.5 - 30.5			
30.5 - 36.5			
36.5 - 42.5			
42.5 - 48.5			

Researcher B

Key Terms

Define the key terms based upon the above example for Researcher A.

Exercise:

Problem: Population

Exercise:

Problem: Sample

Exercise:

Problem: Parameter

Exercise:

Problem: Statistic

Exercise:

Problem: Variable

Exercise:

Problem: Data

Discussion Questions

Discuss the following questions and then answer in complete sentences.

Exercise:

Problem: List two reasons why the data may differ.

Exercise:

Problem:

Can you tell if one researcher is correct and the other one is incorrect?
Why?

Exercise:

Problem: Would you expect the data to be identical? Why or why not?

Exercise:

Problem: How could the researchers gather random data?

Exercise:

Problem:

Suppose that the first researcher conducted his survey by randomly choosing one state in the nation and then randomly picking 40 patients from that state. What sampling method would that researcher have used?

Exercise:

Problem:

Suppose that the second researcher conducted his survey by choosing 40 patients he knew. What sampling method would that researcher have used? What concerns would you have about this data set, based upon the data collection method?

Descriptive Statistics

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Display data graphically and interpret graphs: stemplots, histograms and boxplots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

Introduction

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called "**Descriptive Statistics**". You will learn to calculate, and even more importantly, to interpret these measurements and graphs.

Displaying Data

This module provides a brief introduction into the ways graphs and charts can be used to provide visual representations of data.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample. The graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly.

Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar chart, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), pie charts, and the boxplot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs and bar graphs. Our emphasis will be on histograms and boxplots.

Stem and Leaf Graphs (Stemplots), Line Graphs and Bar Graphs
This module introduces the use of stem-and-leaf graphs (stemplots), line graphs and bar graphs for describing a set of data visually.

One simple graph, the **stem-and-leaf graph** or **stem plot**, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a **final significant digit**. For example, 23 has stem 2 and leaf 3. Four hundred thirty-two (432) has stem 43 and leaf 2. Five thousand four hundred thirty-two (5,432) has stem 543 and leaf 2. The decimal 9.3 has stem 9 and leaf 3. Write the stems in a vertical line from smallest to the largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example: For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest): 3342494953555561636768686969727374788083888888909294949496100	
Stem	Leaf
3	3
4	299
5	355
6	1378899

Stem	Leaf
7	2348
8	03888
9	0244446
10	0

Stem-and-Leaf Diagram

The stem plot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% of the scores were in the 90's or 100, a fairly high number of As.

The stem plot is a quick way to graph and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An **outlier** is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value**. When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers. In the example above, there were no outliers.

Example:

Create a stem plot using the data:

1.11.52.32.52.73.23.33.33.53.84.0 4.24.54.54.74.85.55.66.56.712.3

The data are the distance (in kilometers) from a home to the nearest supermarket.

Exercise:

Problem:

1. Are there any values that might possibly be outliers?
2. Do the data seem to have any concentration of values?

Note: The leaves are to the right of the decimal.

Solution:

The value 12.3 may be an outlier. Values appear to concentrate at 3 and 4 kilometers.

Stem	Leaf
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	

Stem	Leaf
9	
10	
11	
12	3

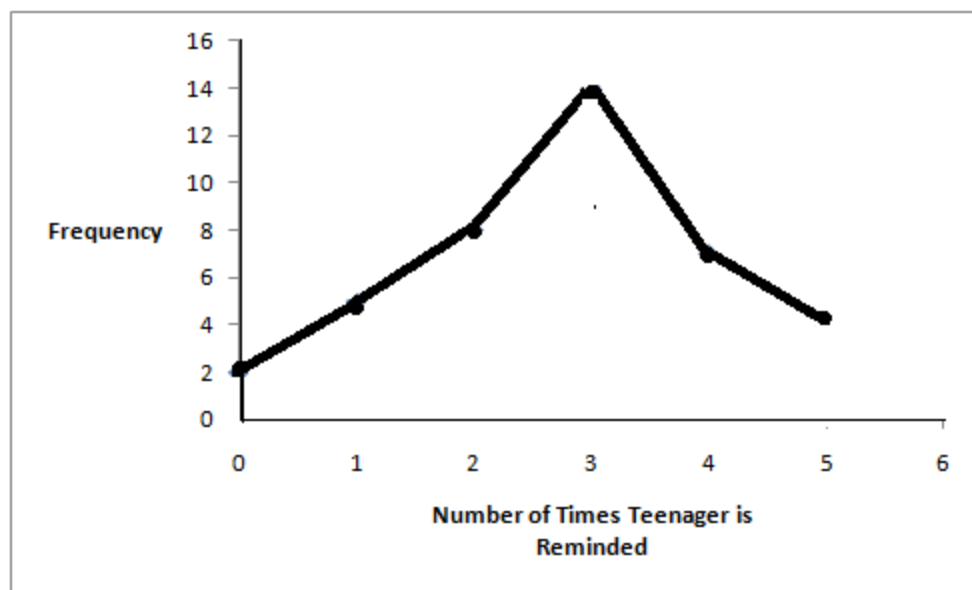
Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in the example, the **x-axis** consists of **data values** and the **y-axis** consists of **frequency points**. The frequency points are connected.

Example:

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his/her chores. The results are shown in the table and the line graph.

Number of times teenager is reminded	Frequency
0	2
1	5

Number of times teenager is reminded	Frequency
2	8
3	14
4	7
5	4



Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes and they can be vertical or horizontal.

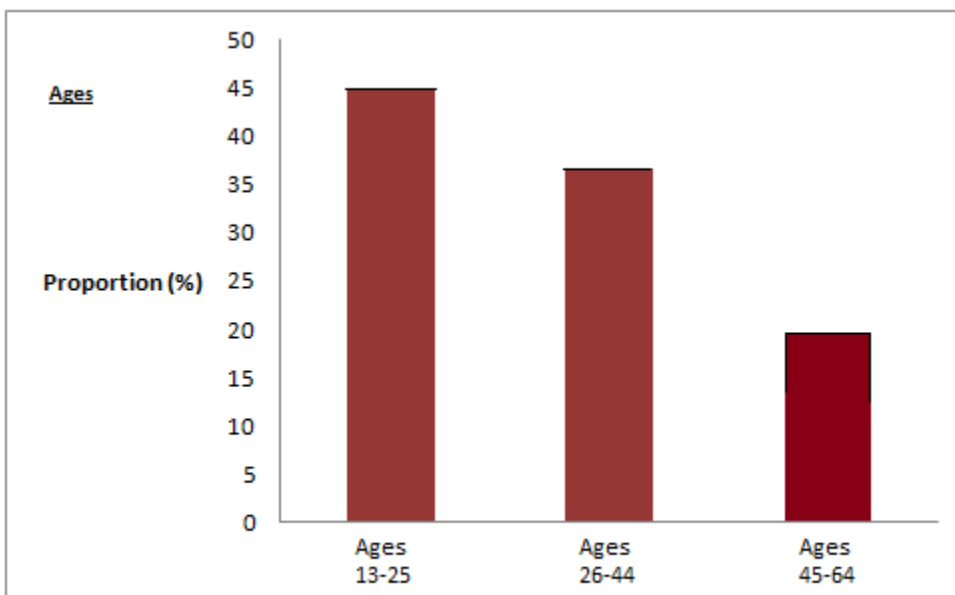
The **bar graph** shown in **Example 4** has age groups represented on the **x-axis** and proportions on the **y-axis**.

Example:

By the end of 2011, in the United States, Facebook had over 146 million users. The table shows three age groups, the number of users in each age group and the proportion (%) of users in each age group. **Source:**

<http://www.kenburbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/>

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13 - 25	65,082,280	45%
26 - 44	53,300,200	36%
45 - 64	27,885,100	19%



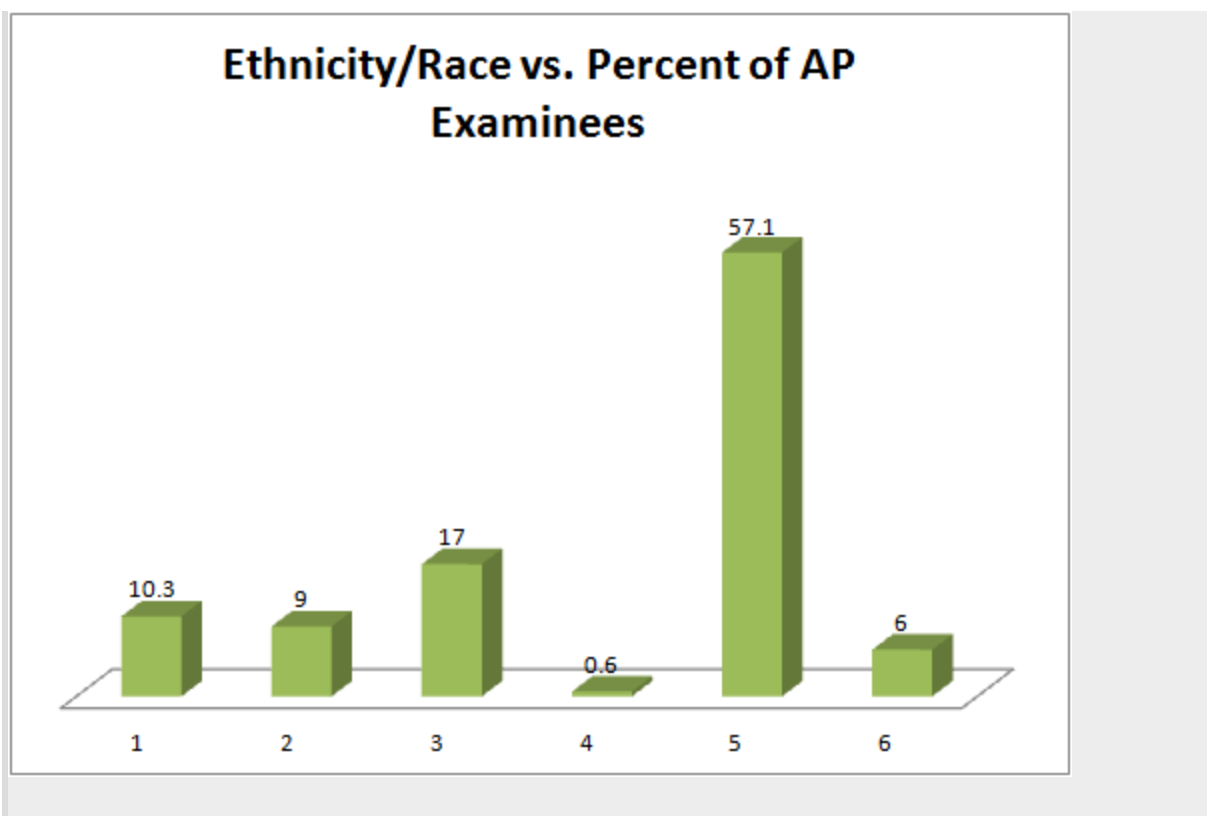
Example:

The columns in the table below contain the race/ethnicity of U.S. Public Schools: High School Class of 2011, percentages for the Advanced Placement Examinee Population for that class and percentages for the Overall Student Population. The 3-dimensional graph shows the Race/Ethnicity of U.S. Public Schools (qualitative data) on the **x-axis** and Advanced Placement Examinee Population percentages on the **y-axis**.

(Source: <http://www.collegeboard.com> and Source:

<http://apreport.collegeboard.org/goals-and-findings/promoting-equity>)

Race/Ethnicity	AP Examinee Population	Overall Student Population
1 = Asian, Asian American or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%



Go to [Outcomes of Education Figure 22](#) for an example of a bar graph that shows unemployment rates of persons 25 years and older for 2009.

Note: This book contains instructions for constructing a **histogram** and a **box plot** for the TI-83+ and TI-84 calculators. You can find additional instructions for using these calculators on the [Texas Instruments \(TI\) website](#).

Glossary

Outlier

An observation that does not fit the rest of the data.

Histograms

This module provides an overview of Descriptive Statistics: Histogram as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A **histogram** consists of contiguous boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either **Frequency** or **relative frequency**. The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data. (The next section tells you how to calculate the center and the spread.)

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. (In the chapter on [Sampling and Data](#), we defined frequency as the number of times an answer occurs.) If:

- f = frequency
- n = total number of data values (or the sum of the individual frequencies), and
- RF = relative frequency,

then:

Equation:

$$\text{RF} = \frac{f}{n}$$

For example, if 3 students in Mr. Ahab's English class of 40 students received from 90% to 100%, then,

$$f = 3, n = 40, \text{ and } RF = \frac{f}{n} = \frac{3}{40} = 0.075$$

Seven and a half percent of the students received 90% to 100%. Ninety percent to 100 % are quantitative measures.

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of from 5 to 15 bars or classes for clarity. Choose a starting point for the first interval to be less than the smallest data value. A **convenient starting point** is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 ($6.1 - 0.05 = 6.05$). We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 ($1.5 - 0.005 = 1.495$). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 ($1.0 - .0005 = 0.9995$). If all the data happen to be integers and the smallest value is 2, then a convenient starting point is 1.5 ($2 - 0.5 = 1.5$). Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary.

Example:

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data since height is measured.

60 60.5 61 61 61.5

63.5 63.5 63.5

64 64 64 64 64 64 64 64.5 64.5 64.5 64.5 64.5 64.5 64.5 64.5

66 66 66 66 66 66 66 66 66 66 66.5 66.5 66.5 66.5 66.5 66.5 66.5 66.5

66.5 66.5 66.5 67 67 67 67 67 67 67 67 67 67 67 67 67 67.5 67.5 67.5 67.5

67.5 67.5 67.5

68 68 69 69 69 69 69 69 69 69 69 69 69.5 69.5 69.5 69.5 69.5

70 70 70 70 70 70 70.5 70.5 70.5 71 71 71

72 72 72 72.5 72.5 73 73.5

74

The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.

$60 - 0.05 = 59.95$ which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.

The largest value is 74. $74 + 0.05 = 74.05$ is the ending value.

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose 8 bars.

Equation:

$$\frac{74.05 - 59.95}{8} = 1.76$$

Note: We will round up to 2 and make each bar or class interval 2 units wide. Rounding up to 2 is one way to prevent a value from falling on a boundary. Rounding to the next number is necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work.

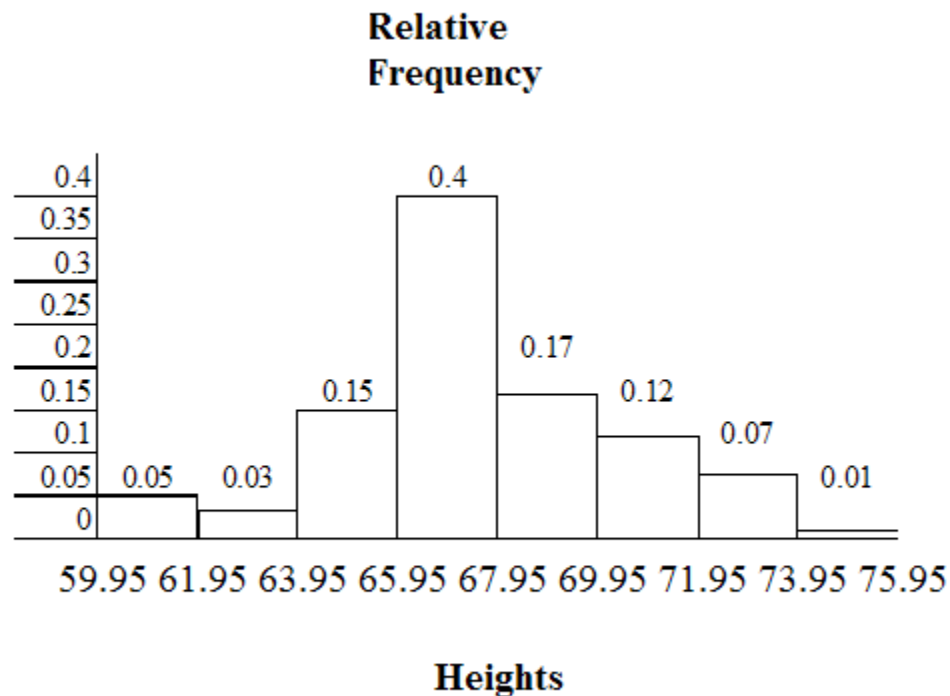
The boundaries are:

- 59.95
- $59.95 + 2 = 61.95$
- $61.95 + 2 = 63.95$
- $63.95 + 2 = 65.95$
- $65.95 + 2 = 67.95$
- $67.95 + 2 = 69.95$
- $69.95 + 2 = 71.95$
- $71.95 + 2 = 73.95$

- $73.95 + 2 = 75.95$

The heights 60 through 61.5 inches are in the interval 59.95 - 61.95. The heights that are 63.5 are in the interval 61.95 - 63.95. The heights that are 64 through 64.5 are in the interval 63.95 - 65.95. The heights 66 through 67.5 are in the interval 65.95 - 67.95. The heights 68 through 69.5 are in the interval 67.95 - 69.95. The heights 70 through 71 are in the interval 69.95 - 71.95. The heights 72 through 73.5 are in the interval 71.95 - 73.95. The height 74 is in the interval 73.95 - 75.95.

The following histogram displays the heights on the x-axis and relative frequency on the y-axis.



Example:

The following data are the number of books bought by 50 part-time college students at ABC College. The number of books is discrete data since books

are counted.

1 1 1 1 1 1 1 1 1 1 1

2 2 2 2 2 2 2 2 2 2

3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3

4 4 4 4 4 4

5 5 5 5 5

6 6

Eleven students buy 1 book. Ten students buy 2 books. Sixteen students buy 3 books. Six students buy 4 books. Five students buy 5 books. Two students buy 6 books.

Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.

Exercise:

Problem:

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6 and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _____ to _____, the 5 in the middle of the interval from _____ to _____, and the _____ in the middle of the interval from _____ to _____.

Solution:

- 3.5 to 4.5
- 4.5 to 5.5
- 6
- 5.5 to 6.5

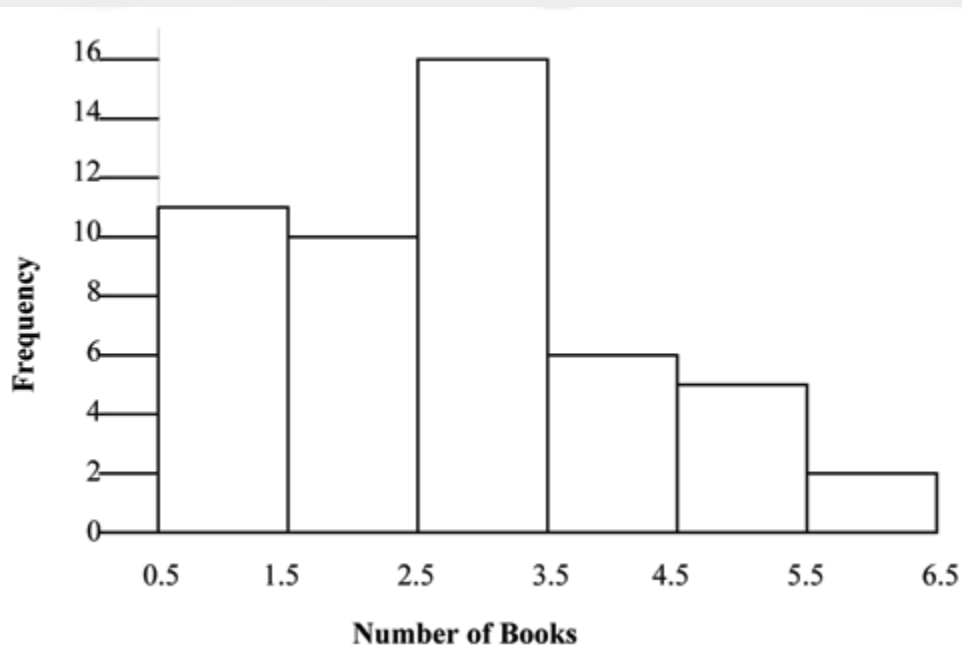
Calculate the number of bars as follows:

Equation:

$$\frac{6.5 - 0.5}{\text{bars}} = 1$$

where 1 is the width of a bar. Therefore, bars = 6.

The following histogram displays the number of books on the x-axis and the frequency on the y-axis.



Using the TI-83, 83+, 84, 84+ Calculator Instructions

Go to the Appendix (14:Appendix) in the menu on the left. There are calculator instructions for entering data and for creating a customized histogram. Create the histogram for Example 2.

- Press Y=. Press CLEAR to clear out any equations.
- Press STAT 1:EDIT. If L1 has data in it, arrow up into the name L1, press CLEAR and arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6
- Into L2, enter 11, 10, 16, 6, 5, 2
- Press WINDOW. Make Xmin = .5, Xmax = 6.5, Xscl = (6.5 - .5)/6, Ymin = -1, Ymax = 20, Yscl = 1, Xres = 1

- Press 2nd Y=. Start by pressing 4:Plotsoff ENTER.
- Press 2nd Y=. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3rd picture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2nd 1). Arrow down to Freq. Enter L2 (2nd 2).
- Press GRAPH
- Use the TRACE key and the arrow keys to examine the histogram.

Optional Collaborative Exercise

Count the money (bills and change) in your pocket or purse. Your instructor will record the amounts. As a class, construct a histogram displaying the data. Discuss how many intervals you think is appropriate. You may want to experiment with the number of intervals. Discuss, also, the shape of the histogram.

Record the data, in dollars (for example, 1.25 dollars).

Construct a histogram.

Glossary

Frequency

The number of times a value of the data occurs.

Relative Frequency

The ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes.

Measures of the Location of the Data

Descriptive Statistics: Measuring the Location of Data explains percentiles and quartiles and is part of the collection col10555 written by Barbara Illowsky and Susan Dean. Roberta Bloom contributed the section "Interpreting Percentiles, Quartile and the Median."

The common measures of location are [quartiles](#) and [percentiles](#) (%iles). Quartiles are special percentiles. The first quartile, Q_1 is the same as the 25th percentile (25th %ile) and the third quartile, Q_3 , is the same as the 75th percentile (75th %ile). The median, M , is called both the second quartile and the 50th percentile (50th %ile).

Note: Quartiles are given special attention in the Box Plots module in this chapter.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Recall that quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The [interquartile range](#) is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

Equation:

$$\text{IQR} = Q_3 - Q_1$$

The IQR can help to determine potential **outliers**. A value is suspected to be a **potential outlier if it is less than $(1.5)(\text{IQR})$ below the first quartile or more than $(1.5)(\text{IQR})$ above the third quartile**. Potential outliers always need further investigation.

Example:

Exercise:

Problem:

For the following 13 real estate prices, calculate the IQR and determine if any prices are outliers. Prices are in dollars. (*Source: San Jose Mercury News*)

389,950 230,500 158,000 479,000 639,000 114,950 5,500,000 387,000
659,000 529,000 575,000 488,800 1,095,000

Solution:

Order the data from smallest to largest.

114,950 158,000 230,500 387,000 389,950 479,000 488,800 529,000
575,000 639,000 659,000 1,095,000 5,500,000

$$M = 488,800$$

$$Q_1 = \frac{230500 + 387000}{2} = 308750$$

$$Q_3 = \frac{639000 + 659000}{2} = 649000$$

$$\text{IQR} = 649000 - 308750 = 340250$$

$$(1.5)(\text{IQR}) = (1.5)(340250) = 510375$$

$$Q_1 - (1.5)(\text{IQR}) = 308750 - 510375 = -201625$$

$$Q_3 + (1.5)(\text{IQR}) = 649000 + 510375 = 1159375$$

No house price is less than -201625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Example:**Exercise:****Problem:**

For the two data sets in the [test scores example](#), find the following:

- **a** The interquartile range. Compare the two interquartile ranges.
- **b** Any outliers in either set.
- **c** The 30th percentile and the 80th percentile for each set. How much data falls below the 30th percentile? Above the 80th percentile?

Solution:

For the IQRs, see the [answer to the test scores example](#). The first data set has the larger IQR, so the scores between Q3 and Q1 (middle 50%) for the first data set are more spread out and not clustered about the median.

First Data Set

- $\left(\frac{3}{2}\right) \cdot (\text{IQR}) = \left(\frac{3}{2}\right) \cdot (26.5) = 39.75$
- $X_{\max} - Q3 = 99 - 82.5 = 16.5$
- $Q1 - X_{\min} = 56 - 32 = 24$

$\left(\frac{3}{2}\right) \cdot (\text{IQR}) = 39.75$ is larger than 16.5 and larger than 24, so the first set has no outliers.

Second Data Set

- $\left(\frac{3}{2}\right) \cdot (\text{IQR}) = \left(\frac{3}{2}\right) \cdot (11) = 16.5$
- $X_{\max} - Q3 = 98 - 89 = 9$
- $Q1 - X_{\min} = 78 - 25.5 = 52.5$

$\left(\frac{3}{2}\right) \cdot (\text{IQR}) = 16.5$ is larger than 9 but smaller than 52.5, so for the second set 45 and 25.5 are outliers.

To find the percentiles, create a frequency, relative frequency, and cumulative relative frequency chart (see ["Frequency" from the Sampling and](#)

[Data Chapter](#)). Get the percentiles from that chart.

First Data Set

- 30th %ile (between the 6th and 7th values) = $\frac{(56 + 59)}{2} = 57.5$
- 80th %ile (between the 16th and 17th values) = $\frac{(84 + 84.5)}{2} = 84.25$

Second Data Set

- 30th %ile (7th value) = 78
- 80th %ile (18th value) = 90

30% of the data falls below the 30th %ile, and 20% falls above the 80th %ile.

Example:

Finding Quartiles and Percentiles Using a Table

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were (student data):

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Find the 28th percentile: Notice the 0.28 in the "cumulative relative frequency" column. 28% of 50 data values = 14. There are 14 values less than the 28th %ile. They include the two 4s, the five 5s, and the seven 6s. The 28th %ile is between the last 6 and the first 7. **The 28th %ile is 6.5.**

Find the median: Look again at the "cumulative relative frequency " column and find 0.52. The median is the 50th %ile or the second quartile. 50% of 50 = 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th %ile is between the 25th (7) and 26th (7) values. **The median is 7.**

Find the third quartile: The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the 4s, 5s, 6s and 7s, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th %ile, then, must be an 8 .** Another way to look at the problem is to find 75% of 50 (= 37.5) and round up to 38. The third quartile, Q_3 , is the 38th value which is an 8. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Example:

Exercise:

Problem: Using the table:

1. Find the 80th percentile.
2. Find the 90th percentile.
3. Find the first quartile.
4. What is another name for the first quartile?

Solution:

1. $\frac{(8+9)}{2} = 8.5$

Look where cum. rel. freq. = 0.80. 80% of the data is 8 or less. 80th %ile is between the last 8 and first 9.

2. 9
3. 6
4. First Quartile = 25th %ile

Collaborative Classroom Exercise: Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions.

1. How many students were surveyed?
2. What kind of sampling did you do?
3. Construct a table of the data.
4. Construct 2 different histograms. For each, starting value = _____ ending value = _____.
5. Use the table to find the median, first quartile, and third quartile.
6. Construct a box plot.
7. Use the table to find the following:
 - The 10th percentile
 - The 70th percentile
 - The percent of students who own less than 4 sweaters

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order, from smallest to largest. $p\%$ of data values are less than or equal to the p th percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad". The interpretation of whether a certain percentile is good or bad depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good"; in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to properly interpret percentiles is important not only when describing data, but is also important in later chapters of this textbook when calculating probabilities.

Guideline:

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information:

- information about the context of the situation being considered,
- the data value (value of the variable) that represents the percentile,
- the percent of individuals or items with data values below the percentile.
- Additionally, you may also choose to state the percent of individuals or items with data values above the percentile.

Example:

On a timed math test, the first quartile for times for finishing the exam was 35 minutes. Interpret the first quartile in the context of this situation.

- 25% of students finished the exam in 35 minutes or less.
- 75% of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Example:

On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

- 70% of students answered 16 or fewer questions correctly.
- 30% of students answered 16 or more questions correctly.
- Note: A high percentile could be considered good, as answering more questions correctly is desirable.

Example:

At a certain community college, it was found that the 30th percentile of credit units that students are enrolled for is 7 units. Interpret the 30th percentile in the context of this situation.

- 30% of students are enrolled in 7 or fewer credit units
- 70% of students are enrolled in 7 or more credit units
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Do the following Practice Problems for Interpreting Percentiles**Exercise:****Problem:**

- **a** For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
- **b** The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.

- **c** A bicyclist in the 90th percentile of a bicycle race between two towns completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

Solution:

- **a** For runners in a race it is more desirable to have a low percentile for finish time. A low percentile means a short time, which is faster.
- **b** INTERPRETATION: 20% of runners finished the race in 5.2 minutes or less. 80% of runners finished the race in 5.2 minutes or longer.
- **c** He is among the slowest cyclists (90% of cyclists were faster than him.) INTERPRETATION: 90% of cyclists had a finish time of 1 hour, 12 minutes or less. Only 10% of cyclists had a finish time of 1 hour, 12 minutes or longer

Exercise:

Problem:

- **a** For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
- **b** The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Solution:

- **a** For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed, which is faster.
- **b** INTERPRETATION: 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

Exercise:

Problem:

On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

Solution:

On an exam you would prefer a high percentile; higher percentiles correspond to higher grades on the exam.

Exercise:**Problem:**

Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

Solution:

When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than you did. In this context, you would prefer a wait time corresponding to a lower percentile. INTERPRETATION: 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

Exercise:**Problem:**

In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

Solution:

Li should be pleased. Her salary is relatively high compared to other recent college grads. 78% of recent college graduates earn less than Li does. 22% of recent college graduates earn more than Li does.

Exercise:

Problem:

In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had \$1700 in damage and was in the 90th percentile. Should the manufacturer and/or a consumer be pleased or upset by this result? Explain. Write a sentence that interprets the 90th percentile in the context of this problem.

Solution:

The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample.

INTERPRETATION: 90% of the crash tested cars had damage repair costs of \$1700 or less; only 10% had damage repair costs of \$1700 or more.

Exercise:**Problem:**

- The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:
 - a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
 - b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percent of students from each high school are "eligible in the local context"?
-

Solution:

- **a** The top 12% of students are those who are at or above the **88th percentile** of admissions index scores.
- **b** The **top 4%** of students' GPAs are at or above the 96th percentile, making the top 4% of students "eligible in the local context".

Exercise:

Problem:

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is \$240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

Solution:

You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost \$240,000 or less. 66% of houses cost \$240,000 or more.

**With contributions from Roberta Bloom

Glossary

Interquartile Range (IRQ)

The distance between the third quartile (Q3) and the first quartile (Q1). $IQR = Q3 - Q1$.

Outlier

An observation that does not fit the rest of the data.

Percentile

A number that divides ordered data into hundredths.

Example:

Let a data set contain 200 ordered observations starting with $\{2.3, 2.7, 2.8, 2.9, 2.9, 3.0, \dots\}$. Then the first percentile is $\frac{(2.7+2.8)}{2} = 2.75$, because 1% of the data is to the left of this point on the number line and 99% of the data is on its right. The second percentile is $\frac{(2.9+2.9)}{2} = 2.9$. Percentiles may or may not be part of the data. In this example, the first percentile is not in the data, but the second percentile is. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Quartiles

The numbers that separate the data into quarters. Quartiles may or may not be part of the data. The second quartile is the median of the data.

Measures of the Center of the Data

This chapter discusses measuring descriptive statistical information using the center of the data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts (previously discussed under box plots in this chapter). The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

Note: The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean" and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

The mean can also be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the sample mean is an x with a bar over it (pronounced " x bar"): \bar{x} .

The Greek letter μ (pronounced "mew") represents the population mean. One of the requirements for the sample mean to be a good estimate of the population mean is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:

11122344444

Equation:

$$x = \frac{1 + 1 + 1 + 2 + 2 + 3 + 4 + 4 + 4 + 4 + 4}{11} = 2.7$$

Equation:

$$x = \frac{3 \times 1 + 2 \times 2 + 1 \times 3 + 5 \times 4}{11} = 2.7$$

In the second calculation for the sample mean, the frequencies are 3, 2, 1, and 5.

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by 2 after the data has been ordered. For example, if the total number of data values is 97, then $\frac{n+1}{2} = \frac{97+1}{2} = 49$. The median is the 49th value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2} = 50.5$. The median occurs midway between the 50th and 51st values. The location of the median and the value of the median are **not** the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example:

Exercise:

Problem:

AIDS data indicating the number of months an AIDS patient lives after taking a new antibody drug are as follows (smallest to largest):

34881011121314151516161717182122222424252626272729293132333
33434353740444447

Calculate the mean and the median.

Solution:

The calculation for the mean is:

$$x = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+\dots+35+37+40+(44)(2)+47]}{40} = 23.6$$

To find the median, **M**, first use the formula for the location. The location is:

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5$$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):

34881011121314151516161717182122222424
25262627272929313233333434353740444447

$$M = \frac{24+24}{2} = 24$$

The median is 24.

Using the TI-83,83+,84, 84+ Calculators

Calculator Instructions are located in the menu item 14:Appendix (Notes for the TI-83, 83+, 84, 84+ Calculators).

- Enter data into the list editor. Press STAT 1:EDIT
- Put the data values in list L1.
- Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and ENTER.
- Press the down and up arrow keys to scroll.

$$x = 23.6, M = 24$$

Example:

Exercise:

Problem:

Suppose that, in a small town of 50 people, one person earns \$5,000,000 per year and the other 49 each earn \$30,000. Which is the better measure of the "center," the mean or the median?

Solution:

$$x = \frac{5000000 + 49 \times 30000}{50} = 129400$$

$$M = 30000$$

(There are 49 people who earn \$30,000 and one person who earns \$5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

Another measure of the center is the mode. The [mode](#) is the most frequent value. If a data set has two values that occur the same number of times, then the set is bimodal.

Example:**Statistics exam scores for 20 students are as follows**

Statistics exam scores for 20 students are as follows:

50 53 59 59 63 63 72 72 72 72 72 76 78 81 83 84 84 84 90 93

Exercise:

Problem: Find the mode.

Solution:

The most frequent score is 72, which occurs five times. Mode = 72.

Example:

Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

Note: The mode can be calculated for qualitative data as well as for quantitative data.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean \bar{x} of the sample is very likely to get closer and closer to μ . This is discussed in more detail in **The Central Limit Theorem**.

Note: The formula for the mean is located in the [Summary of Formulas](#) section course.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a [sampling distribution](#) as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency). Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	Relative Frequency
0	5/30
1	15/30
2	6/30
3	4/30
4	1/30

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A [statistic](#) is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean \bar{x} is an example of a statistic which estimates the population mean μ .

Glossary

Mean

A number that measures the central tendency. A common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by \bar{x}) is

$\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Median

A number that separates ordered data into halves. Half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Mode

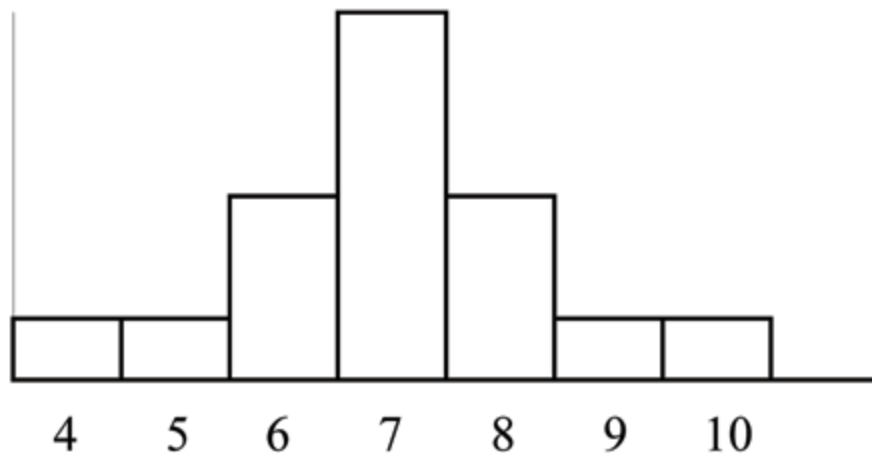
The value that appears most frequently in a set of data.

Skewness and the Mean, Median, and Mode

Consider the following data set:

4 5 6 6 6 7 7 7 7 7 8 8 8 9 10

This data set produces the histogram shown below. Each interval has width one and each value is located in the middle of an interval.

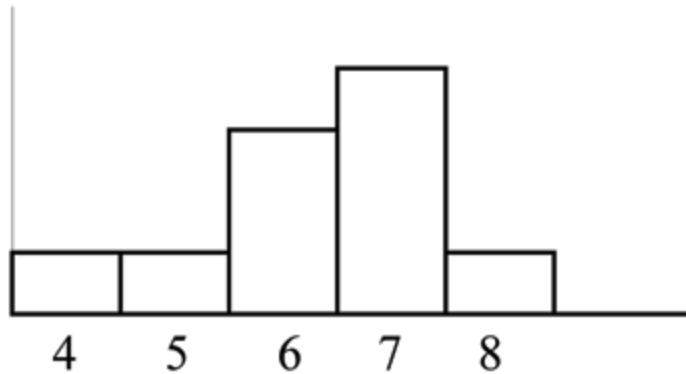


The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each 7 for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal) and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data:

4 5 6 6 6 7 7 7 7 8

is not symmetrical. The right-hand side seems "chopped off" compared to the left side. The shape distribution is called **skewed to the left** because it is pulled out to the left.

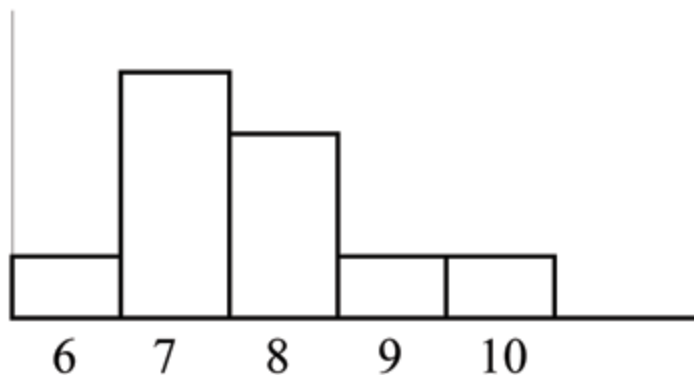


The mean is 6.3, the median is 6.5, and the mode is 7. **Notice that the mean is less than the median and they are both less than the mode.** The mean and the median both reflect the skewing but the mean more so.

The histogram for the data:

6 7 7 7 7 8 8 8 9 10

is also not symmetrical. It is **skewed to the right**.



The mean is 7.7, the median is 7.5, and the mode is 7. Of the three statistics, **the mean is the largest, while the mode is the smallest.** Again, the mean reflects the skewing the most.

To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Understanding and Measuring Variability

This module seeks to explain variability and describe its importance in statistical analysis.

Variability

When we measure people or other subjects the scores usually vary; they are not all the same. Note in the table below that all the students have the same class rank so these data do not tell us much. However, when we use a different type of data we see that the scores are spread out; we can say that they vary.

Class Rank		
Name	Class Rank	Hours Completed
Stanley	Junior	86
Jelani	Junior	54
Christin	Junior	62
Kendra	Junior	60
Kristin	Junior	72
Tykida	Junior	66
Amanda	Junior	77
Brittany	Junior	56

I like to think of explaining variability as the essence of using data to understand psychology. If we all had the same score that score would contain little or no information. However, if each person has a different score then that tells us something. In the table below we can see that Americans differ on the amount of education they have. They also vary in terms of employment status and income. We can look at the way the scores vary and see that, in general, people who have more education are less likely to be unemployed and tend to have higher incomes. One way to describe the variability is the **range** which is simply the difference between the highest value and the lowest value. In the case of income the range is \$1,735 - \$471 or \$1,264.

Earnings and unemployment rates by educational attainment

Education attained	Unemployment rate in 2012 (Percent)	Median weekly earnings
Doctoral degree	2.5	\$1,624.00
Professional degree	2.1	\$1,735.00
Master's degree	3.5	\$1,300.00
Bachelor's degree	4.5	\$1,066.00
Associate's degree	6.2	\$785.00
Some college, no degree	7.7	\$727.00
High school diploma	8.3	\$652.00
Less than a high school diploma	12.4	\$471.00

Source: U.S. Bureau of Labor Statistics, Current Population Survey

These two groups of students have the same mean height but it would be silly to say that, with regards to height, the two groups are the same. What is the difference? They have differing amounts of variability. One set of scores (heights) are spread out more than the other. The **standard deviation** is our favorite measure of variability.



Measures of the Spread of the Data

Descriptive Statistics: Measuring the Spread of Data explains standard deviation as a measure of variation in data and is part of the collection col10555 written by Barbara Illowsky and Susan Dean. Roberta Bloom made contributions that helped to clarify the standard deviation and the variance.

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation.

The [standard deviation](#) is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set
- can be used to determine whether a particular data value is close to or far from the mean

The standard deviation provides a measure of the overall variation in a data set

The standard deviation is always positive or 0. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying waiting times at the checkout line for customers at supermarket A and supermarket B; the average wait time at both markets is 5 minutes. At market A, the standard deviation for the waiting time is 2 minutes; at market B the standard deviation for the waiting time is 4 minutes.

Because market B has a higher standard deviation, we know that there is more variation in the waiting times at market B. Overall, wait times at market B are more spread out from the average; wait times at market A are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that Rosa and Binh both shop at Market A. Rosa waits for 7 minutes and Binh waits for 1 minute at the checkout counter. At market A, the mean wait time is 5 minutes and the standard deviation is 2 minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for 7 minutes:

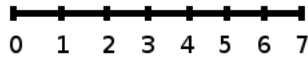
- 7 is 2 minutes longer than the average of 5; 2 minutes is equal to one standard deviation.
- Rosa's wait time of 7 minutes is **2 minutes longer than the average** of 5 minutes.
- Rosa's wait time of 7 minutes is **one standard deviation above the average** of 5 minutes.

Binh waits for 1 minute.

- 1 is 4 minutes less than the average of 5; 4 minutes is equal to two standard deviations.
- Binh's wait time of 1 minute is **4 minutes less than the average** of 5 minutes.
- Binh's wait time of 1 minute is **two standard deviations below the average** of 5 minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than 2 standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than 2 standard deviations. (We will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put 5 and 7 on a number line, 7 is to the right of 5. We say, then, that 7 is **one** standard deviation to the **right** of 5 because $5 + (1)(2) = 7$.

If 1 were also part of the data set, then 1 is **two** standard deviations to the **left** of 5 because $5 + (-2)(2) = 1$.



- In general, a **value** = **mean** + (**#ofSTDEV**)(**standard deviation**)
- where #ofSTDEVs = the number of standard deviations
- 7 is **one standard deviation more than the mean** of 5 because: $7=5+(1)(2)$
- 1 is **two standard deviations less than the mean** of 5 because: $1=5+(-2)(2)$

The equation **value** = **mean** + (**#ofSTDEVs**)(**standard deviation**) can be expressed for a sample and for a population:

- **sample:** $x = \bar{x} + (\text{\#ofSTDEV})(s)$
- **Population:** $x = \mu + (\text{\#ofSTDEV})(\sigma)$

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol \bar{x} is the sample mean and the Greek symbol μ is the population mean.

Calculating the Standard Deviation

If x is a number, then the difference " x - mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - \bar{x}$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ .

To calculate the standard deviation, we need to calculate the variance first. The **variance** is an **average of the squares of the deviations** (the $x - \bar{x}$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by **N**, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by **n-1**, one less than the number of items in the sample. You can see that in the formulas below.

Formulas for the Sample Standard Deviation

- $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$ or $s = \sqrt{\frac{\Sigma f \cdot (x-\bar{x})^2}{n-1}}$
- For the sample standard deviation, the denominator is **n-1**, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

- $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f \cdot (x-\mu)^2}{N}}$
- For the population standard deviation, the denominator is **N**, the number of items in the population.

In these formulas, f represents the frequency with which a value appears. For example, if a value appears once, f is 1. If a value appears three times in the data set or population, f is 3.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in **Descriptive Statistics: Measuring the Center of the Data**. How much the statistic varies from one sample to another is known as the [sampling variability of a statistic](#). You typically measure the sampling variability of a statistic by its standard error. The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in **The Central Limit Theorem** (not now). The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

Note: In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83,83+,84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean.

Example:

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of $n = 20$ fifth grade students. The ages are rounded to the nearest half year:

9 9.5 9.5 10 10 10 10 10.5 10.5 10.5 10.5 11 11 11 11 11 11 11.5 11.5 11.5

Equation:

$$\bar{x} = \frac{9 + 9.5 \times 2 + 10 \times 4 + 10.5 \times 4 + 11 \times 6 + 11.5 \times 3}{20} = 10.525$$

The average age is 10.53 years, rounded to 2 places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s .

Data	Freq.	Deviations	Deviations ²	(Freq.)(Deviations ²)
x	f	$(x - \bar{x})$	$(x - \bar{x})^2$	$(f)(x - \bar{x})^2$
9	1	$9 - 10.525 = -1.525$	$(-1.525)^2 = 2.325625$	$1 \times 2.325625 = 2.325625$
9.5	2	$9.5 - 10.525 = -1.025$	$(-1.025)^2 = 1.050625$	$2 \times 1.050625 = 2.101250$
10	4	$10 - 10.525 = -0.525$	$(-0.525)^2 = 0.275625$	$4 \times .275625 = 1.1025$
10.5	4	$10.5 - 10.525 = -0.025$	$(-0.025)^2 = 0.000625$	$4 \times .000625 = .0025$
11	6	$11 - 10.525 = 0.475$	$(0.475)^2 = 0.225625$	$6 \times .225625 = 1.35375$

Data	Freq.	Deviations	Deviations ²	(Freq.)(Deviations ²)
11.5	3	$11.5 - 10.525 = 0.975$	$(0.975)^2 = 0.950625$	$3 \times .950625 = 2.851875$

The sample variance, s^2 , is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 - 1):

$$s^2 = \frac{9.7375}{20-1} = 0.5125$$

The **sample standard deviation** s is equal to the square root of the sample variance:

$$s = \sqrt{0.5125} = .715891 \text{ Rounded to two decimal places, } s = 0.72$$

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

Exercise:

Problem: Verify the mean and standard deviation calculated above on your calculator or computer.

Solution:

Using the TI-83,83+,84+ Calculators

- Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.
- Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
- Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
- $\bar{x}=10.525$
- Use S_x because this is sample data (not a population): $S_x=0.715891$

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**
- For a sample: $x = \bar{x} + (\text{\#ofSTDEVs})(s)$
- For a population: $x = \mu + (\text{\#ofSTDEVs})(\sigma)$
- For this example, use $x = \bar{x} + (\text{\#ofSTDEVs})(s)$ because the data is from a sample

Exercise:

Problem: Find the value that is 1 standard deviation above the mean. Find $(\bar{x} + 1s)$.

Solution:

$$(\bar{x} + 1s) = 10.53 + (1)(0.72) = 11.25$$

Exercise:

Problem: Find the value that is two standard deviations below the mean. Find $(\bar{x} - 2s)$.

Solution:

$$(\bar{x} - 2s) = 10.53 - (2)(0.72) = 9.09$$

Exercise:

Problem: Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution:

- $(x - 1.5s) = 10.53 - (1.5)(0.72) = 9.45$
- $(x + 1.5s) = 10.53 + (1.5)(0.72) = 11.61$

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11. The deviations 0.97 and 0.47 indicate that. A positive deviation occurs when the data value is greater than the mean. A negative deviation occurs when the data value is less than the mean; the deviation is -1.525 for the data value 9. **If you add the deviations, the sum is always zero.** (For this example, there are $n=20$ deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by $n=20$, the calculation divided by $n-1=20-1=19$ because the data is a sample. For the **sample** variance, we divide by the sample size minus one ($n-1$). Why not divide by n ? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by $(n-1)$ gives a better estimate of the population variance.

Note: Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ , is either zero or larger than zero. When the standard deviation is 0, there is no spread; that is, all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data.**

Note: The formula for the standard deviation is at the end of the chapter.

Example: Exercise:

Problem: Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

3342494953555561 6367686869697273 7478808388888890 929494949496100

- **a** Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
- **b** Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:
 - **i** The sample mean
 - **ii** The sample standard deviation
 - **iii** The median
 - **iv** The first quartile
 - **v** The third quartile
 - **vi** IQR
- **c** Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

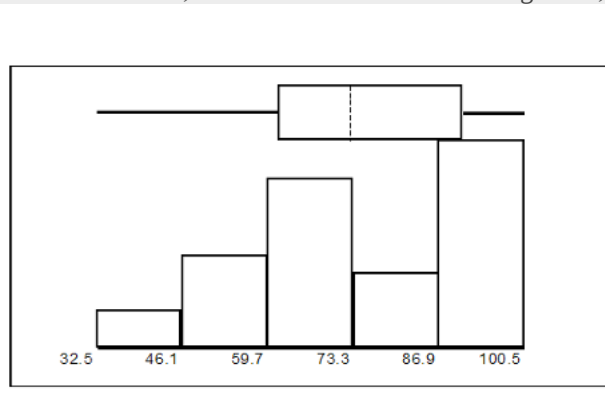
Solution:

- **a**

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

- **b**
 - **i**The sample mean = 73.5
 - **ii**The sample standard deviation = 17.9
 - **iii**The median = 73
 - **iv**The first quartile = 61
 - **v**The third quartile = 90
 - **vi**IQR = 90 - 61 = 29
- **c**The x-axis goes from 32.5 to 100.5; y-axis goes from -2.4 to 15 for the histogram; number of intervals is 5 for the histogram so the width of an interval is (100.5 - 32.5) divided by 5 which is equal to 13.6. Endpoints of the intervals: starting point is 32.5, 32.5+13.6 = 46.1, 46.1+13.6 = 59.7, 59.7+13.6 = 73.3, 73.3+13.6 = 86.9, 86.9+13.6 = 100.5 = the ending value; No data values fall on an interval boundary.



The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater ($73 - 33 = 40$) than the spread in the upper 50% ($100 - 73 = 27$). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores (IQR = 29) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, it can be misleading to compare the data values directly.

- For each data value, calculate how many standard deviations the value is away from its mean.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\#ofSTDEVs = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol z. In symbols, the formulas become:

Sample	$x = \bar{x} + z s$	$z = \frac{x - \bar{x}}{s}$
Population	$x = \mu + z \sigma$	$z = \frac{x - \mu}{\sigma}$

Example:

Exercise:

Problem:

Two students, John and Ali, from different high schools, wanted to find out who had the highest G.P.A. when compared to his school. Which student had the highest G.P.A. when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Solution:

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$$\#ofSTDEVs = \frac{\text{value} - \text{mean}}{\text{standard deviation}} ; z = \frac{x - \mu}{\sigma}$$

$$\text{For John, } z = \#ofSTDEVs = \frac{2.85 - 3.0}{0.7} = -0.21$$

$$\text{For Ali, } z = \#ofSTDEVs = \frac{77 - 80}{10} = -0.3$$

John has the better G.P.A. when compared to his school because his G.P.A. is 0.21 standard deviations **below** his school's mean while Ali's G.P.A. is 0.3 standard deviations **below** his school's mean.

John's z-score of -0.21 is higher than Ali's z-score of -0.3 . For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within 2 standard deviations of the mean.
- At least 89% of the data is within 3 standard deviations of the mean.
- At least 95% of the data is within 4 1/2 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is MOUND-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within 1 standard deviation of the mean.
- Approximately 95% of the data is within 2 standard deviations of the mean.
- More than 99% of the data is within 3 standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is mound-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

**With contributions from Roberta Bloom

Glossary

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Variance

Mean of the squared deviations from the mean. Square of the standard deviation. For a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and \bar{x} is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and 1.

Summary of Formulas

A summary of useful formulas used in examining descriptive statistics

Commonly Used Symbols

- The symbol Σ means to add or to find the sum.
- n = the number of data values in a sample
- N = the number of people, things, etc. in the population
- \bar{x} = the sample mean
- s = the sample standard deviation
- μ = the population mean
- σ = the population standard deviation
- f = frequency
- x = numerical value

Commonly Used Expressions

- $x \cdot f$ = A value multiplied by its respective frequency
- $\sum x$ = The sum of the values
- $\sum x \cdot f$ = The sum of values multiplied by their respective frequencies
- $(x - \bar{x})$ or $(x - \mu)$ = Deviations from the mean (how far a value is from the mean)
- $(x - \bar{x})^2$ or $(x - \mu)^2$ = Deviations squared
- $f(x - \bar{x})^2$ or $f(x - \mu)^2$ = The deviations squared and multiplied by their frequencies

Mean Formulas:

- $\bar{x} = \frac{\sum x}{n}$ or $\bar{x} = \frac{\sum f \cdot x}{n}$
- $\mu = \frac{\sum x}{N}$ or $\mu = \frac{\sum f \cdot x}{N}$

Standard Deviation Formulas:

- $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$ or $s = \sqrt{\frac{\sum f \cdot (x - \bar{x})^2}{n-1}}$
- $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f \cdot (x - \mu)^2}{N}}$

Formulas Relating a Value, the Mean, and the Standard Deviation:

- value = mean + (#ofSTDEVs)(standard deviation), where #ofSTDEVs = the number of standard deviations
- $x = \bar{x} + (\text{\#ofSTDEVs})(s)$
- $x = \mu + (\text{\#ofSTDEVs})(\sigma)$

Normal Distribution: Introduction

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.

Introduction

The normal, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real estate prices fit a normal distribution. The normal distribution is extremely important but it cannot be applied to everything in the real world.

In this chapter, you will study the normal distribution, the standard normal, and applications associated with them.

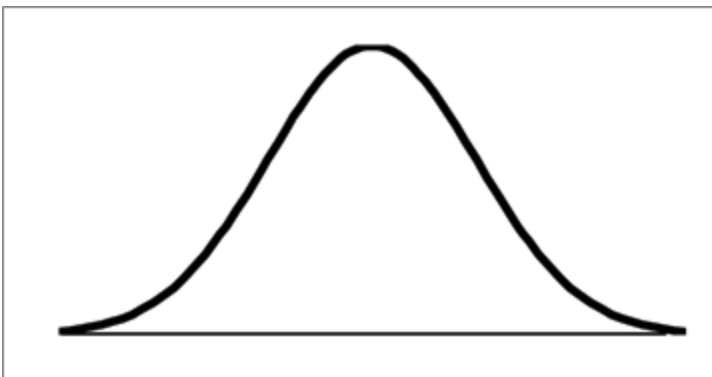
Optional Collaborative Classroom Activity

Your instructor will record the heights of both men and women in your class, separately. Draw histograms of your data. Then draw a smooth curve through each histogram. Is each curve somewhat bell-shaped? Do you think that if you had recorded 200 data values for men and 200 for women that the curves would look bell-shaped? Calculate the mean for each data set. Write the means on the x-axis of the appropriate graph below the peak.

Shade the approximate area that represents the probability that one randomly chosen male is taller than 72 inches. Shade the approximate area that represents the probability that one randomly chosen female is shorter than 60 inches. If the total area under each curve is one, does either probability appear to be more than 0.5?

The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and the standard deviation (σ), we designate this by writing

NORMAL: $X \sim N(\mu, \sigma)$



The probability density function is a rather complicated function. **Do not memorize it.** It is not necessary.

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma}\right)^2}$$

The cumulative distribution function is $P(X < x)$. It is calculated either by a calculator or a computer or it is looked up in a table. You may use technology when calculating probabilities, but on exams you will need to know how to use a normal distribution like the one below. The full table can be found at the end of the book in the appendix.

z-Score Chart

Use this chart to find the area under a normal curve when finding
an approximation for a binomial distribution.

Negative z-score - value is to the left of the mean.

Positive z-score - value is to the right of the mean.

Negative z-scores:										
Z	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.0
-3.4	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005
-3.2	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007
-3.1	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010
-3.0	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013
-2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019
-2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026

The curve is symmetrical about a vertical line drawn through the mean, μ . In theory, the mean is the same as the median since the graph is symmetric about μ . As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ , causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ . A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

Glossary

Normal Distribution

A continuous random variable (RV) with pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \text{ where } \mu \text{ is the mean of the distribution and}$$

σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Variables

Independent and dependent variables

Variables are properties or characteristics of some event, object, or person that can take on different values or amounts (as opposed to constants such as p which do not vary). When conducting research, experimenters often manipulate variables. For example, an experimenter might compare the effectiveness of four types of antidepressants. In this case, the variable is the "type of antidepressant".

Independent variable

When a variable is manipulated by an experimenter

Dependent variable

The experiment seeks to determine the effect of the independent variable on relief from depression. In this example, relief from depression is called a dependent variable.

In general the independent variable is manipulated by the experimenter and its effects on the dependent variable are measured.

Example:

Can blueberries slow down aging?

A study indicates that antioxidants found in blueberries may slow down the process of aging. In this study, 19-month old rats (equivalent to 60-year old humans) were fed either their standard diet or a diet supplemented by either blueberry, strawberry, or spinach powder. After eight weeks, the rats were given memory and motor tests. Although all supplemented rats showed improvement, those supplemented with blueberry powder showed the most notable improvement.

1. What is the independent variable? (diet: blueberries or no blueberries)
2. What are the dependent variables? (memory test and motor skills test)

[More information on the blueberry study.](#)

Example:

Does [beta-carotene](#) protect against cancer?

Beta-carotene supplements have been thought to protect against cancer. However, a study published in the Journal of the National Cancer Institute suggests this is false. The study was conducted with 39,000 women aged 45 and up. These women were randomly assigned to receive a beta-carotene supplement or a [placebo](#), and their health was studied over their lifetime. Cancer rates for women taking the beta-carotene supplement did not differ systematically from the cancer rates of those women taking the placebo.

1. What is the independent variable? (supplements: beta-carotene or placebo)
2. What is the dependent variable? (occurrence of cancer)

Example:

How bright is right?

An automobile manufacturer wants to know how bright brake lights should be in order to minimize the time required for the driver of a following car to realize that the car in front is stopping and to hit the brakes.

1. What is the independent variable? (brightness of brake light)
2. What is the dependent variable? (time to hit brake)

Levels of an Independent Variable

If an experiment compares an experimental treatment with a control treatment, then the independent variable (type of treatment) has two levels: experimental and control. If an experiment were comparing five types of

diets, then the independent variable (type of diet) would have 5 levels. In general, the number of levels of an independent variable is the number of experimental conditions.

Qualitative and Quantitative Variables

An important distinction between variables is between [qualitative](#) and [quantitative](#) variables.

Qualitative variable

Variables that express a qualitative attribute

Example:

Some examples of qualitative variables are hair color, eye color, religion, favorite movie, gender, and so on.

The values of a qualitative variable do not imply a numerical ordering. Values of the variable "religion" differ qualitatively; no ordering of religions is implied. Qualitative variables are sometimes referred to as **categorical variables**. Values on qualitative variables do not imply order, they are simply categories.

Quantitative variables

Variables that are measured in terms of numbers.

Example:

Some examples of quantitative variables are height, weight, and shoe size.

In the study on the effect of diet discussed [above](#), the independent variable was type of supplement: none, strawberry, blueberry, and spinach. The

variable "type of supplement" is a qualitative variable; there is nothing quantitative about it. In contrast, the dependent variable "memory test" is a quantitative variable since memory performance was measured on a quantitative scale (number correct).

Discrete and Continuous Variables

Variables such as number of children in a household are called [discrete variables](#).

Discrete variable

Variable with possible scores of discrete points on the scale.

Example:

A household could have three children or six children, but not 4.53 children.

Other variables such as "time to respond to a question" are [continuous variables](#).

Continuous variable

Variable where the scale is continuous and not made up of discrete steps.

Example:

The response time could be 1.64 seconds, or it could be 1.64237123922121 seconds.

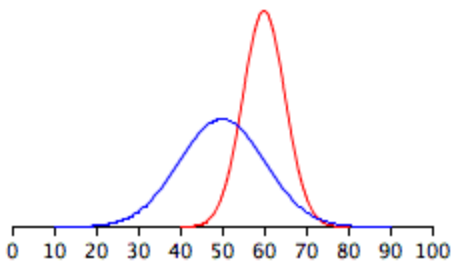
Of course, the practicalities of measurement preclude most measured variables from being truly continuous.

Introduction to Normal Distributions

The normal distribution is the most important and widely used distribution in statistics. It is sometimes called the **bell curve** although the tonal qualities of such a bell would be less than pleasing. It is also called the **Gaussian curve** after the mathematician Karl-Friedrich Gauss. As you will see in the section on the history of the normal distribution, although Gauss played an important role in its history, de Moivre first discovered the normal distribution.

Strictly speaking, it is not correct to talk about **the normal distribution** since there are many normal distributions. Normal distributions can differ in their means and in their standard deviations. [\[link\]](#) shows two normal distributions. The blue distribution has a mean of 50 and a standard deviation of 10; the distribution in red has a mean of 60 and a standard deviation of 5. Both distributions are symmetric with relatively more values at the center of the distribution and relatively few in the tails.

Varieties of Normal Distributions



Normal distributions
differing in mean and
standard deviation.

The density of the normal distribution (the height for a given value on the x axis) of the normal distribution is shown below ([\[link\]](#)). The parameters μ and σ are the mean and standard deviation respectively and define the normal distribution. The symbol e is the base of the natural logarithm and π is the constant pi.

Equation:

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Since this is a non-mathematical treatment of statistics, do not worry if this expression confuses you. We will **not** be referring back to it in later sections.

Six features of normal distributions are listed below. These features are illustrated in more detail in the remaining sections of this chapter.

1. Normal distributions are symmetric around their mean.
2. The mean, median, and mode of a normal distribution are equal.
3. The area under the normal curve is equal to 1.0.
4. Normal distributions are denser in the center and less dense in the tails.
5. Normal distributions are defined by two parameters, the mean (m) and the standard deviation (s).
6. 68% of the area of a normal distribution is within one standard deviation of the mean

Z-scores

If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the z-score is:

Equation:

$$z = \frac{x - \mu}{\sigma}$$

The z-score tells you how many standard deviations that the value x is above (to the right of) or below (to the left of) the mean, μ . Values of x that are larger than the mean have positive z-scores and values of x that are smaller than the mean have negative z-scores. If x equals the mean, then x has a z-score of 0.

Example:

Suppose $X \sim N(5, 6)$. This says that X is a normally distributed random variable with mean $\mu = 5$ and standard deviation $\sigma = 6$. Suppose $x = 17$. Then:

Equation:

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2$$

This means that $x = 17$ is **2 standard deviations** (2σ) above or to the right of the mean $\mu = 5$. The standard deviation is $\sigma = 6$.

Notice that:

Equation:

$$5 + 2 \cdot 6 = 17 \quad (\text{The pattern is } \mu + z\sigma = x.)$$

Now suppose $x=1$. Then:

Equation:

$$z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{6} = -0.67 \quad (\text{rounded to two decimal places})$$

This means that $x = 1$ is 0.67 standard deviations (-0.67σ) below or to the left of the mean $\mu = 5$. Notice that:

$5 + (-0.67)(6)$ is approximately equal to 1 (This has the pattern $\mu + (-0.67)\sigma = 1$)

Summarizing, when z is positive, x is above or to the right of μ and when z is negative, x is to the left of or below μ .

Example:

Some doctors believe that a person can lose 5 pounds, on the average, in a month by reducing his/her fat intake and by exercising consistently.

Suppose weight loss has a normal distribution. Let X = the amount of weight lost (in pounds) by a person in a month. Use a standard deviation of 2 pounds. $X \sim N(5, 2)$. Fill in the blanks.

Exercise:

Problem:

Suppose a person **lost** 10 pounds in a month. The z-score when $x = 10$ pounds is $z = 2.5$ (verify). This z-score tells you that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean _____. (What is the mean?).

Solution:

This z-score tells you that $x = 10$ is **2.5** standard deviations to the **right** of the mean 5.

Exercise:

Problem:

Suppose a person **gained** 3 pounds (a negative weight loss). Then $z =$ _____. This z-score tells you that $x = -3$ is _____ standard deviations to the _____ (right or left) of the mean.

Solution:

$z = -4$. This z-score tells you that $x = -3$ is 4 standard deviations to the **left** of the mean.

Suppose the random variables X and Y have the following normal distributions: $X \sim N(5, 6)$ and $Y \sim N(2, 1)$. If $x = 17$, then $z = 2$. (This was previously shown.) If $y = 4$, what is z ?

Equation:

$$z = \frac{y - \mu}{\sigma} = \frac{4 - 2}{1} = 2 \quad \text{where } \mu=2 \text{ and } \sigma=1.$$

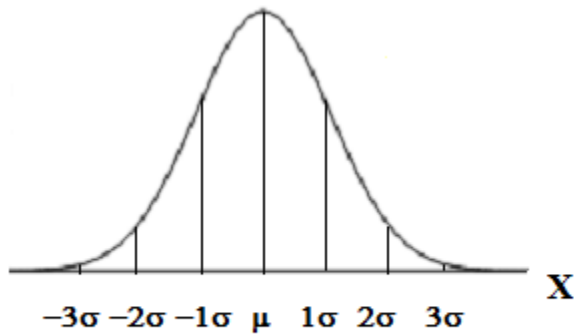
The z-score for $y = 4$ is $z = 2$. This means that 4 is $z = 2$ standard deviations to the right of the mean. Therefore, $x = 17$ and $y = 4$ are both 2 (of **their**) standard deviations to the right of **their** respective means.

The z-score allows us to compare data that are scaled differently. To understand the concept, suppose $X \sim N(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a 6 week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each 2 standard deviations to the right of their means, they represent the same weight gain **relative to their means**.

The Empirical Rule

If X is a random variable and has a normal distribution with mean μ and standard deviation σ then the **Empirical Rule** says (See the figure below)

- About 68.27% of the x values lie between -1σ and $+1\sigma$ of the mean μ (within 1 standard deviation of the mean).
- About 95.45% of the x values lie between -2σ and $+2\sigma$ of the mean μ (within 2 standard deviations of the mean).
- About 99.73% of the x values lie between -3σ and $+3\sigma$ of the mean μ (within 3 standard deviations of the mean). Notice that almost all the x values lie within 3 standard deviations of the mean.
- The z-scores for $+1\sigma$ and -1σ are +1 and -1, respectively.
- The z-scores for $+2\sigma$ and -2σ are +2 and -2, respectively.
- The z-scores for $+3\sigma$ and -3σ are +3 and -3 respectively.



The Empirical Rule is also known as the 68-95-99.7 Rule.

Example:

Suppose X has a normal distribution with mean 50 and standard deviation 6.

- About 68.27% of the x values lie between $-1\sigma = (-1)(6) = -6$ and $1\sigma = (1)(6) = 6$ of the mean 50. The values $50 - 6 = 44$ and $50 + 6 = 56$ are within 1 standard deviation of the mean 50. The z-scores are -1 and +1 for 44 and 56, respectively.
- About 95.45% of the x values lie between $-2\sigma = (-2)(6) = -12$ and $2\sigma = (2)(6) = 12$ of the mean 50. The values $50 - 12 = 38$ and $50 + 12 = 62$ are within 2 standard deviations of the mean 50. The z-scores are -2 and 2 for 38 and 62, respectively.
- About 99.73% of the x values lie between $-3\sigma = (-3)(6) = -18$ and $3\sigma = (3)(6) = 18$ of the mean 50. The values $50 - 18 = 32$ and $50 + 18 = 68$ are within 3 standard deviations of the mean 50. The z-scores are -3 and +3 for 32 and 68, respectively.

The Standard Normal Distribution

The **standard normal distribution** is a normal distribution of **standardized values called z-scores**. A z-score is measured in units of the standard deviation. For example, if the mean of a normal distribution is 5 and the standard deviation is 2, the value 11 is 3 standard deviations above (or to the right of) the mean. The calculation is:

Equation:

$$x = \mu + (z)\sigma = 5 + (3)(2) = 11$$

The z-score is 3.

The mean for the standard normal distribution is 0 and the standard deviation is 1. The transformation

$z = \frac{x-\mu}{\sigma}$ produces the distribution $Z \sim N(0, 1)$. The value x comes from a normal distribution with mean μ and standard deviation σ .

Glossary

Standard Normal Distribution

A continuous random variable (RV) $X \sim N(0,1)$. When X follows the standard normal distribution, it is often noted as $Z \sim N(0,1)$.

z-score

The linear transformation of the form $z = \frac{x-\mu}{\sigma}$. If this transformation is applied to any normal distribution $X \sim N(\mu, \sigma)$, the result is the standard normal distribution $Z \sim N(0,1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ , the result is called the z-score of x . Z-scores allow us to compare data that are normally distributed but scaled differently.

The Normal Curve

Explains the pattern of scores in a normal distribution.

The Normal Curve

A normal curve follows a mathematical pattern and has some important features:

- It is Symmetrical

- Mean=Mode=Median

- It Can be defined in terms of the mean and standard deviation

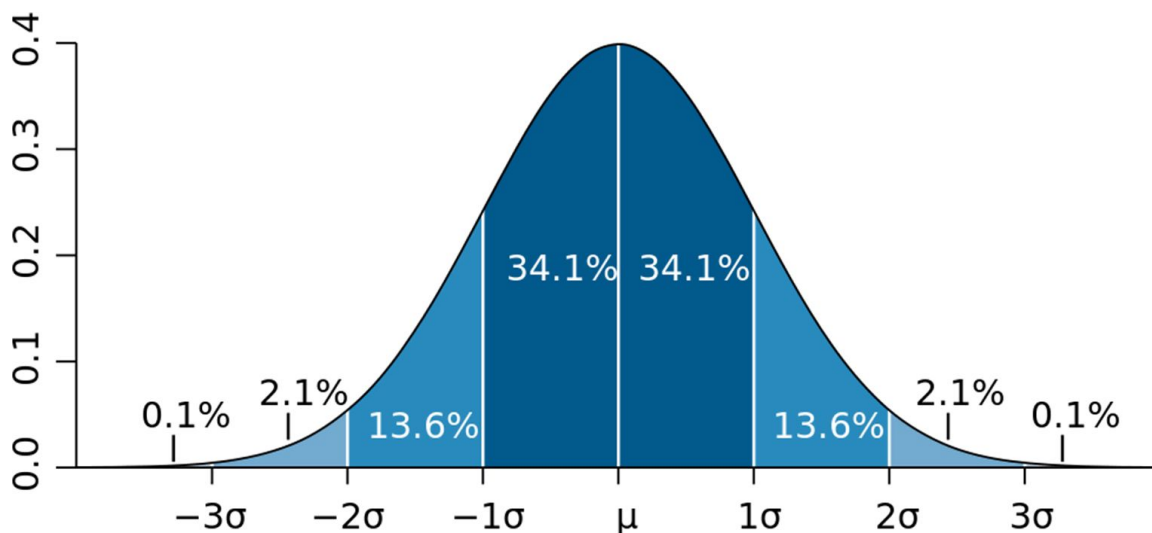
As shown in the diagram below there is a predictable pattern to the distribution of scores.

This is often simplified as the 68-95-99.7 rule. It tells us that :

- 68% of scores lie within one standard deviation of the mean

- 95% lie within two standard deviations of the mean

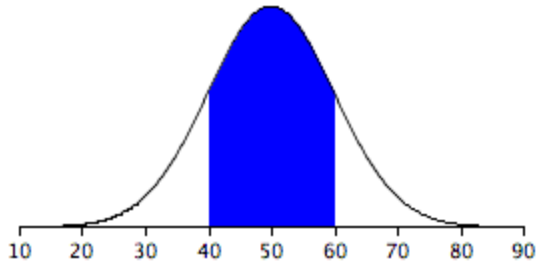
- 99.7% lie within three standard deviations of the mean.



Standard deviation diagram, based an original graph by Jeremy Kemp, in
2005-02-09 [<http://pbeirne.com/Programming/gaussian.ps>].

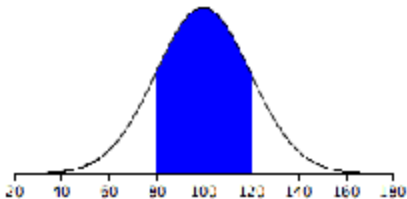
Areas of Normal Distributions

Areas under portions of a normal distribution can be computed by using calculus. Since this is a non-mathematical treatment of statistics, we will rely on computer programs and tables to determine these areas. [\[link\]](#) shows a normal distribution with a mean of 50 and a standard deviation of 10. The shaded area between 40 and 60 contains 68% of the distribution.



Normal distribution with a mean of 50 and standard deviation of 10. 68% of the area is within one standard deviation (10) of the mean (50).

[\[link\]](#) shows a normal distribution with a mean of 100 and a standard deviation of 20. As in Figure 1, 68% of the distribution is within one standard deviation of the mean.



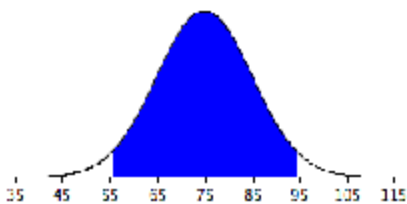
Normal distribution with a mean of 100 and standard deviation of 20. 68% of the area is within

one standard
deviation (20) of the
mean (100).

The normal distributions shown in [\[link\]](#) and [\[link\]](#) are specific examples of the general rule that 68% of the area of any normal distribution is within one standard deviation of the mean.

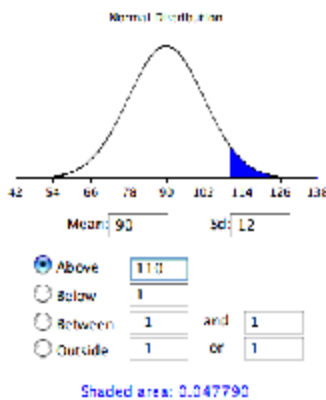
Note: 68% of the area of any normal distribution is within one standard deviation of the mean

[\[link\]](#) shows a normal distribution with a mean of 75 and a standard deviation of 10. The shaded area contains 95% of the area and extends from 55.4 to 94.6. For all normal distributions, 95% of the area is within 1.96 standard deviations of the mean. For quick approximations, it is sometimes useful to round off and use 2 rather than 1.96 as the number of standard deviations you need to extend from the mean so as to include 95% of the area.



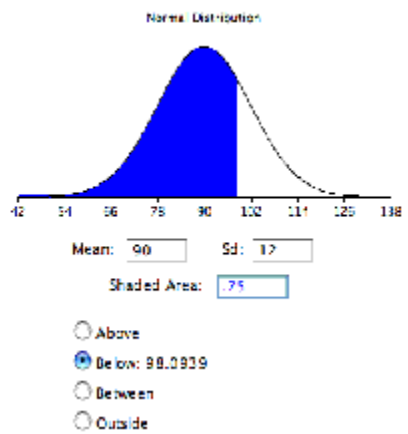
A normal distribution
with a mean of 75 and
a standard deviation
of 10. 95% of the area
is within 1.96
standard deviations of
the mean.

The Java applet "Calculate Area for a given X " can be used to calculate areas under the normal distribution. Use it to find the proportion of a normal distribution with a mean of 90 and a standard deviation of 12 that is above 110. Set the mean to 90 and the standard deviation to 12. Then enter "110" in the box to the right of the radio button "Above." At the bottom of the display you will see that the shaded area is 0.04779. See if you can use the applet to find that the area between 115 and 120 is 0.012401.



Display from
applet showing
the area above
110.

The applet "Calculate X for a given Area" works in reverse. For example, say you wanted to find the score corresponding to the 75th percentile of a normal distribution with a mean of 90 and a standard deviation of 12. You enter 90 for the mean and 12 for the standard deviation. Then, enter 0.75 for the shaded area and click the "Below" button. The area below 98.0939 is 0.75.



Display from applet
showing that the
75th percentile is
98.093.

Calculations of Probabilities

Probabilities are calculated by using technology. There are instructions in the chapter for the TI-83+ and TI-84 calculators.

Note:In the Table of Contents for **Collaborative Statistics**, entry **15. Tables** has a link to a table of normal probabilities. Use the probability tables if so desired, instead of a calculator. The tables include instructions for how to use them.

Example:

If the area to the left is 0.0228, then the area to the right is $1 - 0.0228 = 0.9772$.

Example:

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of 5.

Exercise:

Problem:

Find the probability that a randomly selected student scored more than 65 on the exam.

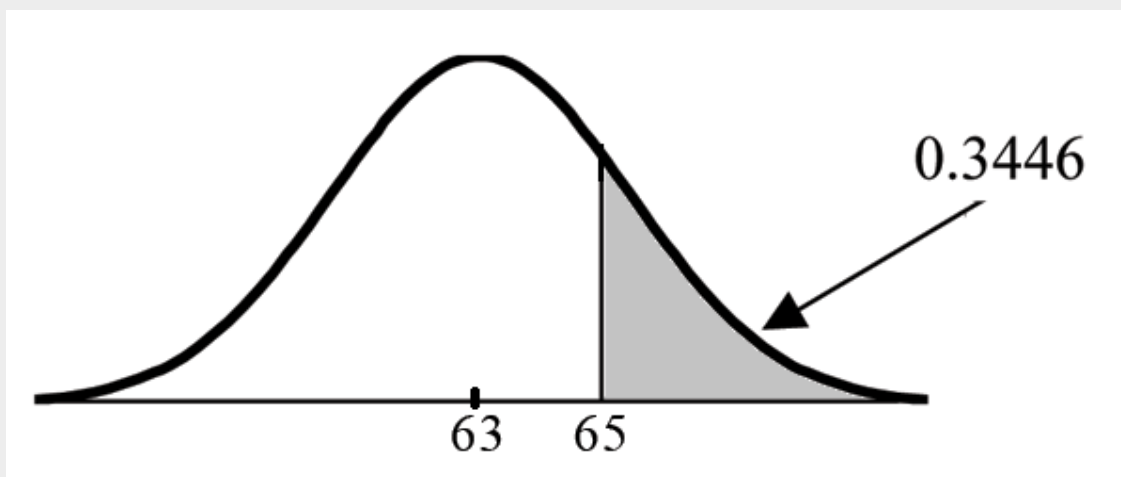
Solution:

Let X = a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$

Draw a graph.

Then, find $P(x > 65)$.

$$P(x > 65) = 0.3446 \text{ (calculator or computer)}$$



The probability that one student scores more than 65 is 0.3446.

Using the TI-83+ or the TI-84 calculators, the calculation is as follows. Go into **2nd DISTR**.

After pressing **2nd DISTR**, press **2:normalcdf**.

The syntax for the instructions are shown below.

`normalcdf(lower value, upper value, mean, standard deviation)` For this problem: `normalcdf(65,1E99,63,5) = 0.3446`. You get 1E99 (= 10^{99}) by pressing **1**, the **EE** key (a 2nd key) and then **99**. Or, you can enter **10^99** instead. The number 10^{99} is way out in the right tail of the normal curve. We are calculating the area between 65 and 10^{99} . In some instances, the lower number of the area might be -1E99 (= -10^{99}). The number -10^{99} is way out in the left tail of the normal curve.

Note: The TI probability program calculates a z-score and then the probability from the z-score. Before technology, the z-score was looked up in a standard normal probability table (because the math involved is too cumbersome) to find the probability. In this example,

a standard normal table with area to the left of the z-score was used. You calculate the z-score and look up the area to the left. The probability is the area to the right.

$$z = \frac{65-63}{5} = 0.4 \quad . \text{Area to the left is } 0.6554.$$
$$P(x > 65) = P(z > 0.4) = 1 - 0.6554 = 0.3446$$

Exercise:

Problem:

Find the probability that a randomly selected student scored less than 85.

Solution:

Draw a graph.

Then find $P(x < 85)$. Shade the graph. $P(x < 85) = 1$ (calculator or computer)

The probability that one student scores less than 85 is approximately 1 (or 100%).

The TI-instructions and answer are as follows:

$$\text{normalcdf}(0,85,63,5) = 1 \text{ (rounds to 1)}$$

Exercise:

Problem:

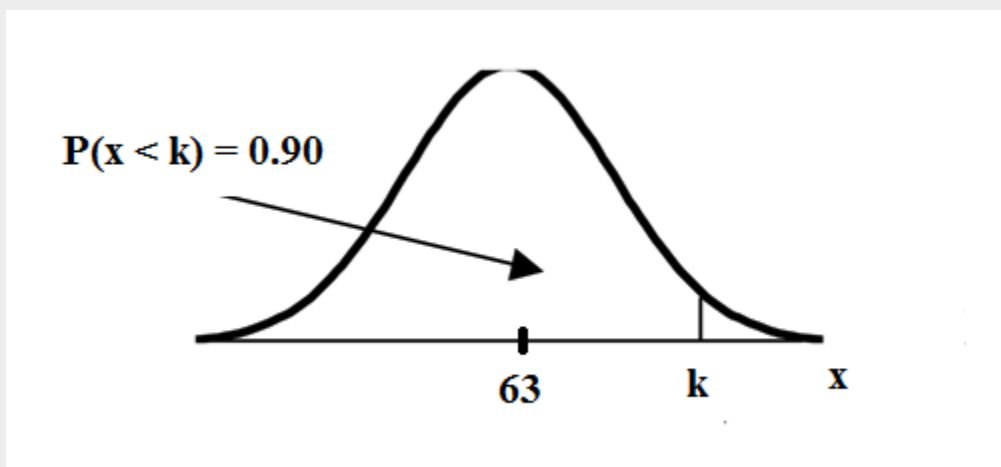
Find the 90th percentile (that is, find the score k that has 90 % of the scores below k and 10% of the scores above k).

Solution:

Find the 90th percentile. For each problem or part of a problem, draw a new graph. Draw the x-axis. Shade the area that corresponds to the 90th percentile.

Let k = the 90th percentile. k is located on the x-axis. $P(x < k)$ is the area to the left of k . The 90th percentile k separates the exam scores into those that are the same or lower than k and those that are the same or higher. Ninety percent of the test scores are the same or lower than k and 10% are the same or higher. k is often called a **critical value**.

$k = 69.4$ (calculator or computer)



The 90th percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above. For the TI-83+ or TI-84 calculators, use **invNorm** in **2nd DISTR**. invNorm(area to the left, mean, standard deviation) For this problem, invNorm(0.90,63,5) = 69.4

Exercise:

Problem:

Find the 70th percentile (that is, find the score k such that 70% of scores are below k and 30% of the scores are above k).

Solution:

Find the 70th percentile.

Draw a new graph and label it appropriately. $k = 65.6$

The 70th percentile is 65.6. This means that 70% of the test scores fall at or below 65.6 and 30% fall at or above.

$$\text{invNorm}(0.70, 63, 5) = 65.6$$

Example:

A computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is 2 hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

Exercise:

Problem:

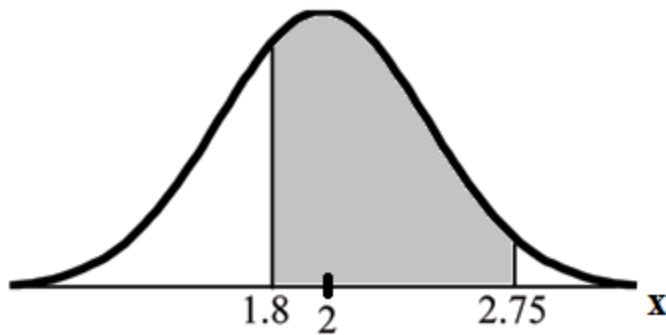
Find the probability that a household personal computer is used between 1.8 and 2.75 hours per day.

Solution:

Let X = the amount of time (in hours) a household personal computer is used for entertainment. $x \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.

Find $P(1.8 < x < 2.75)$.

The probability for which you are looking is the area **between** $x = 1.8$ and $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$



$$\text{normalcdf}(1.8, 2.75, 2, 0.5) = 0.5886$$

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

Exercise:

Problem:

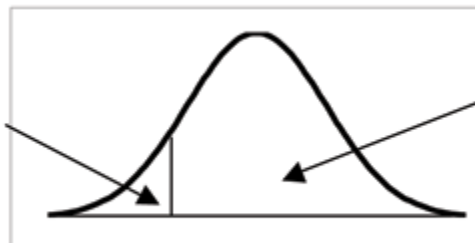
Find the maximum number of hours per day that the bottom quartile of households use a personal computer for entertainment.

Solution:

To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile, k** , where $P(x < k) = 0.25$.

$$k = 1.67$$

$$P(x < k) = 0.25$$



$$P(x > k) = 0.75$$

$$\text{invNorm}(0.25, 2, .5) = 1.66$$

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

Normal Distribution: Calculations of Probabilities

Probabilities are calculated by using technology and the normal distribution table in the appendix of the text.

Example:

If the area to the left is 0.0228, then the area to the right is $1 - 0.0228 = 0.9772$.

Example:

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of 5.

Exercise:**Problem:**

Find the probability that a randomly selected student scored more than 65 on the exam.

Solution:

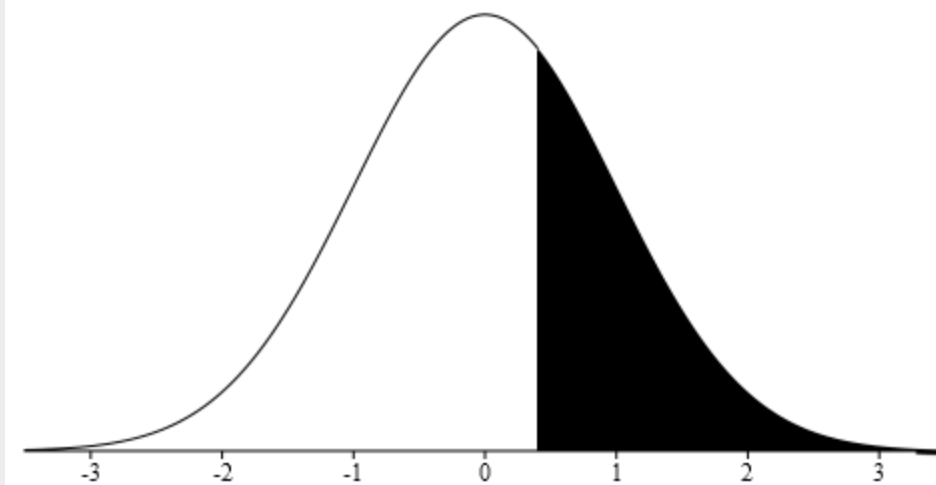
Start by defining the variable and describing the population. Let X = a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$.

Next calculate the z-score, $z = \frac{x - \mu}{\sigma} = \frac{65 - 63}{5} = \frac{2}{5} = 0.40$.

Then draw the normal distribution curve labeling the x-axis with z-scores. Remember that the peak of the distribution will be at zero. Graph where a z-score of +0.40 will be found. We want the probability of being more than this z-score.

$$P(x > 65).$$

$$P(z > +0.40)$$



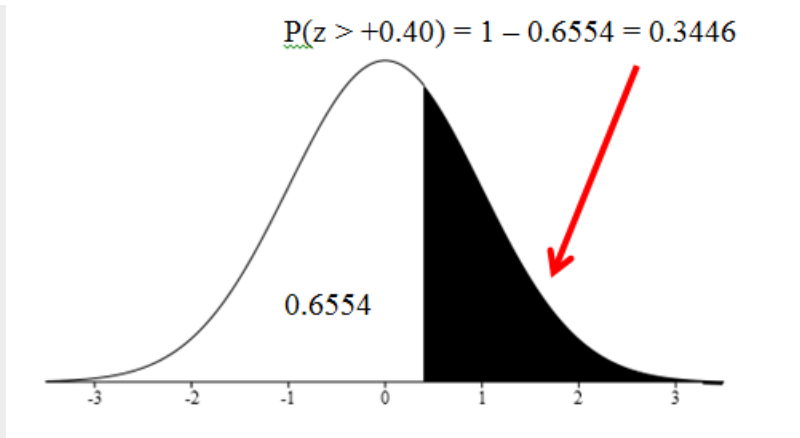
Using the normal distribution table locate the z-score you calculated. The ones and tenths place is found along the side of the table and the hundredths place is located at the top of the table. Our z-score is +0.40 so we will use the positive z-score table and find 0.4 on the side and 0.00 on the top. Our probability is found where this row and column intersect.

↓

Z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157

→

Remember that we are looking for the probability of more than +0.40, $P(x > 65) = P(z > +0.40)$. Our table tells us the probability of being less than our z-score so we need to subtract the probability on the chart from one.



Exercise:

Problem:

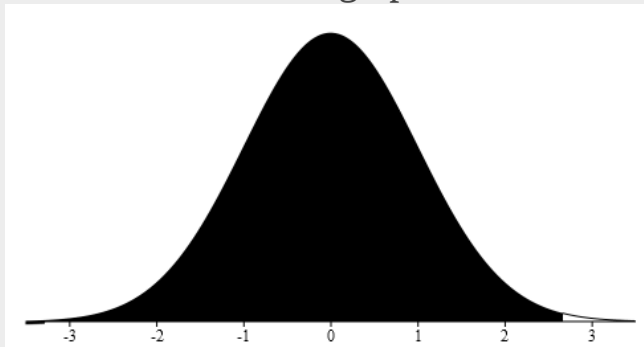
Find the probability that a randomly selected student scored less than 76.3.

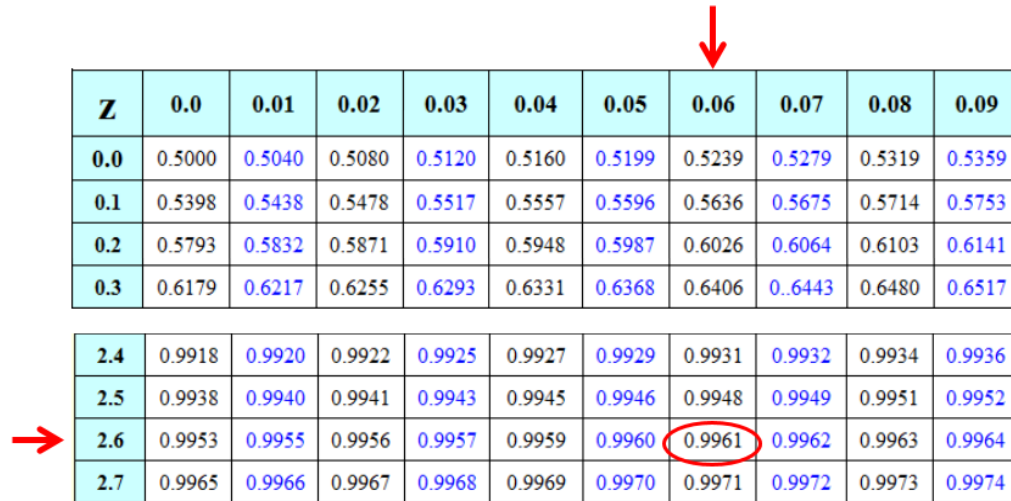
Solution:

Using the same population mean and standard deviation in problem 1, let X = score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$.

Next calculate the z-score, $z = \frac{x - \mu}{\sigma} = \frac{76.3 - 63}{5} = \frac{13.3}{5} = 2.66$

Then draw the normal distribution curve labeling the x-axis with z-scores. Remember that the peak of the distribution will be at zero. Graph where a z-score of +2.66 will be found. We want the probability of being less than this z-score, $P(x < 76.3) = P(z < +2.66)$, so we will shade the graph to the left.





Z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974

Our table tells us the probability of being less than our z-score the probability we see on the table is the answer to our question;
 $P(x < 76.3) = P(z < +2.66) = 0.9961$ or 99.61%.

The probability that one student scores less than 76.3 is approximately .9961 (or 99.61%).

Exercise:

Problem:

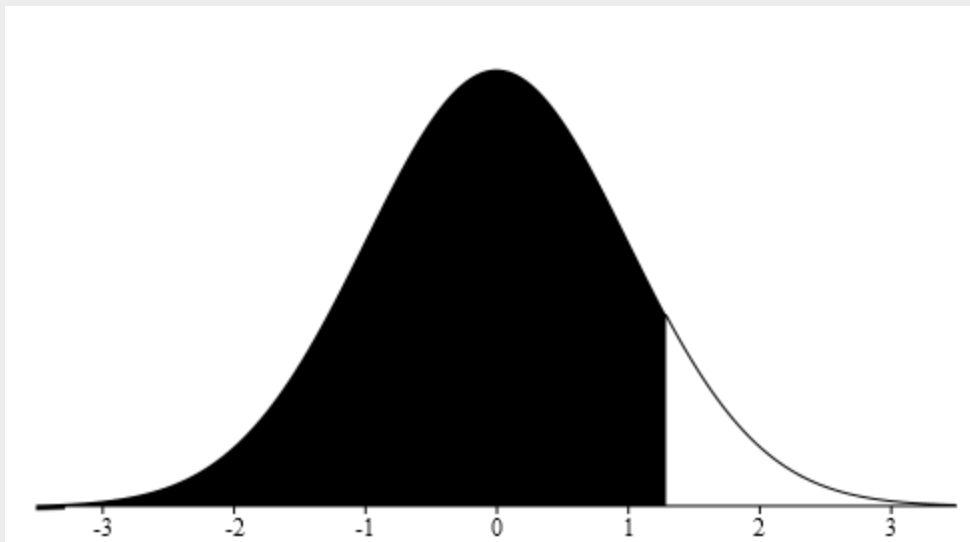
Find the 90th percentile (that is, find the score k that has 90 % of the scores below k and 10% of the scores above k). We are using the same population as in the previous problems, $X \sim N(63, 5)$.

Solution:

Start by using the standard normal table and finding which z-score gives a probability closest to 90% or 0.90. To do this look at the probabilities in the interior of the table, 0.8997 is the closest .090. The z-score for this probability is 1.28.

Z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177

Next draw the graph and shading the area that corresponds to the 90th percentile.



Find the z-score 1.28 on graph and shade the area to the right. Using your z-score, mean and standard deviation put the values you know into the z-score formula. $z = \frac{k-\mu}{\sigma}$ Use algebra to solve for k, the score closest to 90% $1.28 = \frac{k-63}{5}$

The score closest to 90% is 69.4.

Example:

A computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is 2 hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

Exercise:

Problem:

Find the probability that a household personal computer is used between 1.8 and 2.75 hours per day.

Solution:

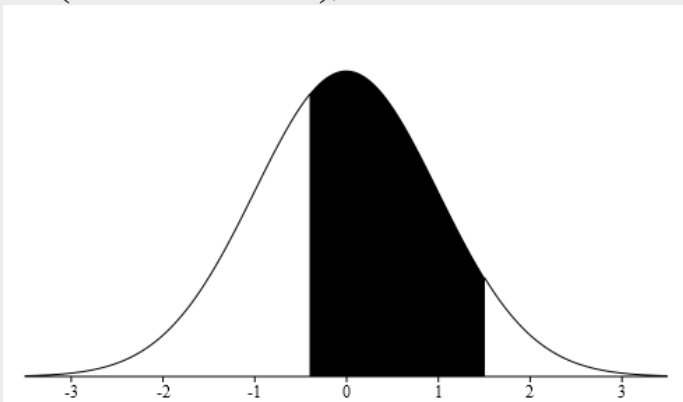
Let X = the amount of time (in hours) a household personal computer is used for entertainment. $x \sim N(2, 0.5)$ where $\mu = 2$ and $\sigma = 0.5$.

Find $P(1.8 < x < 2.75)$. We want to know the probability of being between 1.8 and 2.75 hours. Start by calculating the z-score for each of these values. To calculate the z-score,

$$z = \frac{x - \mu}{\sigma} = \frac{1.8 - 2}{0.5} = \frac{-0.2}{0.5} = -0.40$$

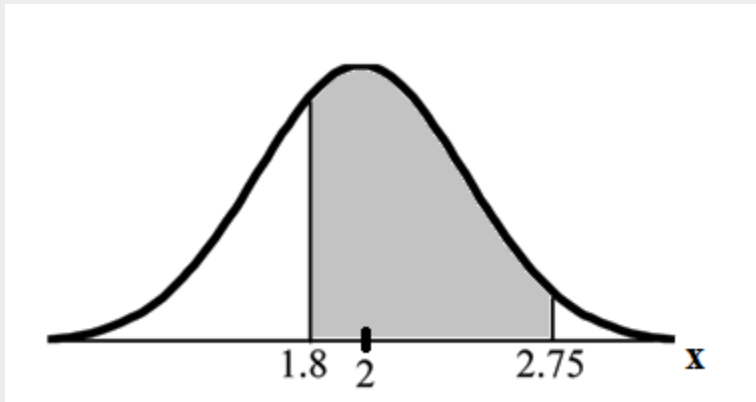
$$z = \frac{x - \mu}{\sigma} = \frac{2.75 - 2}{0.5} = \frac{.75}{0.5} = 1.5$$

Draw the standard normal distribution curve labeling the x-axis with z-scores. Graph where a z-score of -0.40 and +1.50 will be found. We want the probability of being between these z-scores, $P(1.8 < x < 2.75) = P(-0.40 < z < +1.50)$, so we will shade between these two z-scores.



Now look up the two z-scores on the Z-score table to find their probabilities. $P(z < -0.40) = 0.3446$ and $P(z < 1.50) = 0.9332$. To find the probability of being between two values subtract the two probabilities, the larger probability take away the smaller probability, $0.9332 - 0.3446 = 0.5886$.

$$P(-0.40 < z < +1.50) = 0.5886$$



The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

Exercise:

Problem:

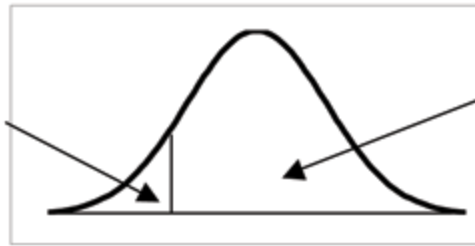
Find the maximum number of hours per day that the bottom quartile of households use a personal computer for entertainment.

Solution:

To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile, k** , where $P(x < k) = 0.25$.

$$k = 1.67$$

$$P(x < k) = 0.25$$



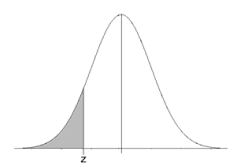
Find the z-score that has a probability closest to 25% or .25. The $P(z < -0.67) = 0.2514$. Using the same population as in the previous problem, $X \sim N(2, 0.5)$ put what you know into the z-score formula and solve for k . $z = \frac{k - \mu}{\sigma}$ Use algebra to solve for k , the score closest to 25% $-0.67 = \frac{k - 2}{0.5}$

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

z Scores with Critical Values
z-table with critical values

Z Score Table

Chart value corresponds to area below z score as shown in the figure below.

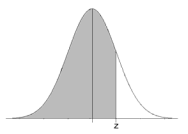


z	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
– 3.4	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
– 3.3	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005
– 3.2	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007
– 3.1	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009
– 3.0	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013
– 2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018
– 2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025
– 2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034
– 2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045
– 2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060
– 2.4	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080

z	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
– 2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104
– 2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136
– 2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174
– 2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222
– 1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281
– 1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351
– 1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436
– 1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537
– 1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655
– 1.4	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793
– 1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951
– 1.2	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131
– 1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335
– 1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562
– 0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814
– 0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090
– 0.7	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389
– 0.6	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709

z	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
– 0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050
– 0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409
– 0.3	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783
– 0.2	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168
– 0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562
– 0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960

Chart value corresponds to area below the positive z score as shown in the figure below.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997

z-Score Critical Values

z-score	-2.576	-1.645	0.675	0.842	1.036	1.282	1.645	2.326	2.576
Probability	.5%	5%	75%	80%	85%	90%	95%	99%	99.5%

The Central Limit Theorem

This module provides a brief introduction to the Central Limit Theorem.

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Recognize the Central Limit Theorem problems.
- Classify continuous word problems by their distributions.
- Apply and interpret the Central Limit Theorem for Means.
- Apply and interpret the Central Limit Theorem for Sums.

Introduction

Why are we so concerned with means? Two reasons are that they give us a middle ground for comparison and they are easy to calculate. In this chapter, you will study means and the Central Limit Theorem.

The Central Limit Theorem (CLT for short) is one of the most powerful and useful ideas in all of statistics. Both alternatives are concerned with drawing finite samples of size n from a population with a known mean, μ , and a known standard deviation, σ . The first alternative says that if we collect samples of size n and n is "large enough," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape. The second alternative says that if we again collect samples of size n that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

In either case, it does not matter what the distribution of the original population is, or whether you even need to know it. The important fact is that the sample means and the sums tend to follow the normal distribution. And, the rest you will learn in this chapter.

The size of the sample, n , that is required in order to be to be 'large enough' depends on the original population from which the samples are drawn. If the original population is far from normal then more observations are

needed for the sample means or the sample sums to be normal. **Sampling is done with replacement.**

Optional Collaborative Classroom Activity

Do the following example in class: Suppose 8 of you roll 1 fair die 10 times, 7 of you roll 2 fair dice 10 times, 9 of you roll 5 fair dice 10 times, and 11 of you roll 10 fair dice 10 times.

Each time a person rolls more than one die, he/she calculates the sample mean of the faces showing. For example, one person might roll 5 fair dice and get a 2, 2, 3, 4, 6 on one roll.

The mean is $\frac{2+2+3+4+6}{5} = 3.4$. The 3.4 is one mean when 5 fair dice are rolled. This same person would roll the 5 dice 9 more times and calculate 9 more means for a total of 10 means.

Your instructor will pass out the dice to several people as described above. Roll your dice 10 times. For each roll, record the faces and find the mean. Round to the nearest 0.5.

Your instructor (and possibly you) will produce one graph (it might be a histogram) for 1 die, one graph for 2 dice, one graph for 5 dice, and one graph for 10 dice. Since the "mean" when you roll one die, is just the face on the die, what distribution do these **means** appear to be representing?

Draw the graph for the means using 2 dice. Do the sample means show any kind of pattern?

Draw the graph for the means using 5 dice. Do you see any pattern emerging?

Finally, draw the graph for the means using 10 dice. Do you see any pattern to the graph? What can you conclude as you increase the number of dice?

As the number of dice rolled increases from 1 to 2 to 5 to 10, the following is happening:

1. The mean of the sample means remains approximately the same.
2. The spread of the sample means (the standard deviation of the sample means) gets smaller.
3. The graph appears steeper and thinner.

You have just demonstrated the Central Limit Theorem (CLT).

The Central Limit Theorem tells you that as you increase the number of dice, **the sample means tend toward a normal distribution (the sampling distribution).**

Glossary

Average

A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ . We are sampling with size n and we are interested in two new RVs - the sample mean, \bar{X} , and the sample sum, ΣX . If the size n of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(n\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

The Central Limit Theorem for Sample Means (Averages)

Suppose X is a random variable with a distribution that may be known or unknown (it can be any distribution). Using a subscript that matches the random variable, suppose:

- μ_X = the mean of X
- σ_X = the standard deviation of X

If you draw random samples of size n , then as n increases, the random variable X which consists of sample means, tends to be **normally distributed** and

$$X \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$$

The Central Limit Theorem for Sample Means says that if you keep drawing larger and larger samples (like rolling 1, 2, 5, and, finally, 10 dice) and **calculating their means** the sample means form their own **normal distribution** (the sampling distribution). The normal distribution has the same mean as the original distribution and a variance that equals the original variance divided by n , the sample size. n is the number of values that are averaged together not the number of times the experiment is done.

To put it more formally, if you draw random samples of size n , the distribution of the random variable X , which consists of sample means, is called the **sampling distribution of the mean**. The sampling distribution of the mean approaches a normal distribution as n , the sample size, increases.

The random variable X has a different z-score associated with it than the random variable X . x is the value of X in one sample.

Equation:

$$z = \frac{x - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}$$

μ_X is both the average of X and of \bar{X} .

$\sigma_X = \frac{\sigma_x}{\sqrt{n}}$ = standard deviation of \bar{X} and is called the standard error of the mean.

Example:

An unknown distribution has a mean of 90 and a standard deviation of 15. Samples of size $n = 25$ are drawn randomly from the population.

Exercise:

Problem:

Find the probability that the **sample mean** is between 85 and 92.

Solution:

Let X = one value from the original unknown population. The probability question asks you to find a probability for the **sample mean**.

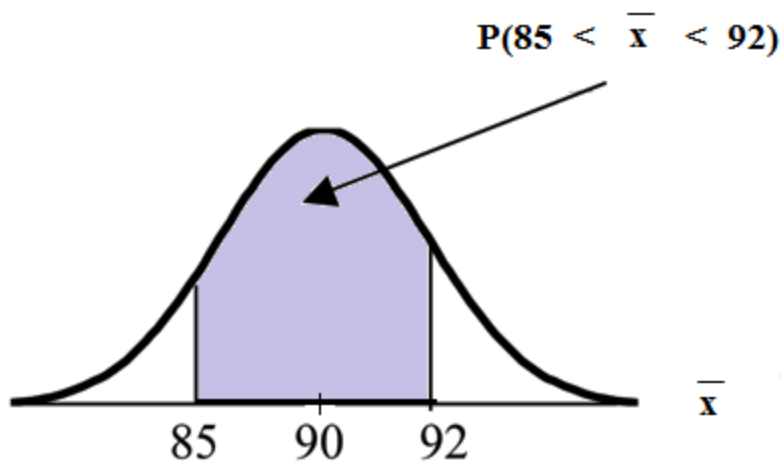
Let \bar{X} = the mean of a sample of size 25. Since $\mu_X = 90$, $\sigma_X = 15$, and $n = 25$;

$$\text{then } \bar{X} \sim N\left(90, \frac{15}{\sqrt{25}}\right)$$

Find $P(85 < \bar{x} < 92)$ Draw a graph.

$$P(85 < \bar{x} < 92) = 0.6997$$

The probability that the sample mean is between 85 and 92 is 0.6997.



TI-83 or 84: `normalcdf`(lower value, upper value, mean, standard error of the mean)

The parameter list is abbreviated (lower value, upper value, μ , $\frac{\sigma}{\sqrt{n}}$)

$$\text{normalcdf}(85, 92, 90, \frac{15}{\sqrt{25}}) = 0.6997$$

Exercise:

Problem:

Find the value that is 2 standard deviations above the expected value (it is 90) of the sample mean.

Solution:

To find the value that is 2 standard deviations above the expected value 90, use the formula

$$\text{value} = \mu_X + (\# \text{ of STDEVs}) \left(\frac{\sigma_X}{\sqrt{n}} \right)$$

$$\text{value} = 90 + 2 \cdot \frac{15}{\sqrt{25}} = 96$$

So, the value that is 2 standard deviations above the expected value is 96.

Example:

The length of time, in hours, it takes an "over 40" group of people to play one soccer match is normally distributed with a **mean of 2 hours** and a **standard deviation of 0.5 hours**. A **sample of size $n = 50$** is drawn randomly from the population.

Exercise:

Problem:

Find the probability that the **sample mean** is between 1.8 hours and 2.3 hours.

Solution:

Let X = the time, in hours, it takes to play one soccer match.

The probability question asks you to find a probability for the **sample mean time, in hours**, it takes to play one soccer match.

Let X = the mean time, in hours, it takes to play one soccer match.

If $\mu_X = \underline{\hspace{2cm}}$, $\sigma_X = \underline{\hspace{2cm}}$, and $n = \underline{\hspace{2cm}}$, then $X \sim N(\underline{\hspace{2cm}}, \underline{\hspace{2cm}})$ by the Central Limit Theorem for Means.

$$\mu_X = 2, \sigma_X = 0.5, n = 50, \text{ and } X \sim N\left(2, \frac{0.5}{\sqrt{50}}\right)$$

Find $P(1.8 < x < 2.3)$. Draw a graph.

$$P(1.8 < x < 2.3) = 0.9977$$

$$\text{normalcdf}(1.8, 2.3, 2, \frac{.5}{\sqrt{50}}) = 0.9977$$

The probability that the mean time is between 1.8 hours and 2.3 hours is _____.

Glossary

Average

A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ . We are sampling with size n and we are interested in two new RVs - the sample mean, \bar{X} , and the sample sum, ΣX . If the size n of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(n\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Normal Distribution

A continuous random variable (RV) with pdf

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, where μ is the mean of the distribution and

σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Error of the Mean

The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$.

The Central Limit Theorem for Sums

Suppose X is a random variable with a distribution that may be **known or unknown** (it can be any distribution) and suppose:

- **a** μ_X = the mean of X
- **b** σ_X = the standard deviation of X

If you draw random samples of size n , then as n increases, the random variable ΣX which consists of sums tends to be **normally distributed** and

$$\Sigma X \sim N(n \cdot \mu_X, \sqrt{n} \cdot \sigma_X)$$

The Central Limit Theorem for Sums says that if you keep drawing larger and larger samples and taking their sums, the sums form their own normal distribution (the sampling distribution) which approaches a normal distribution as the sample size increases. **The normal distribution has a mean equal to the original mean multiplied by the sample size and a standard deviation equal to the original standard deviation multiplied by the square root of the sample size.**

The random variable ΣX has the following z-score associated with it:

- **a** Σx is one sum.
- **b** $z = \frac{\Sigma x - n \cdot \mu_X}{\sqrt{n} \cdot \sigma_X}$
- **a** $n \cdot \mu_X$ = the mean of ΣX
- **b** $\sqrt{n} \cdot \sigma_X$ = standard deviation of ΣX

Example:

An unknown distribution has a mean of 90 and a standard deviation of 15. A sample of size 80 is drawn randomly from the population.

Exercise:

Problem:

- **a**Find the probability that the sum of the 80 values (or the total of the 80 values) is more than 7500.
- **b**Find the sum that is 1.5 standard deviations above the mean of the sums.

Solution:

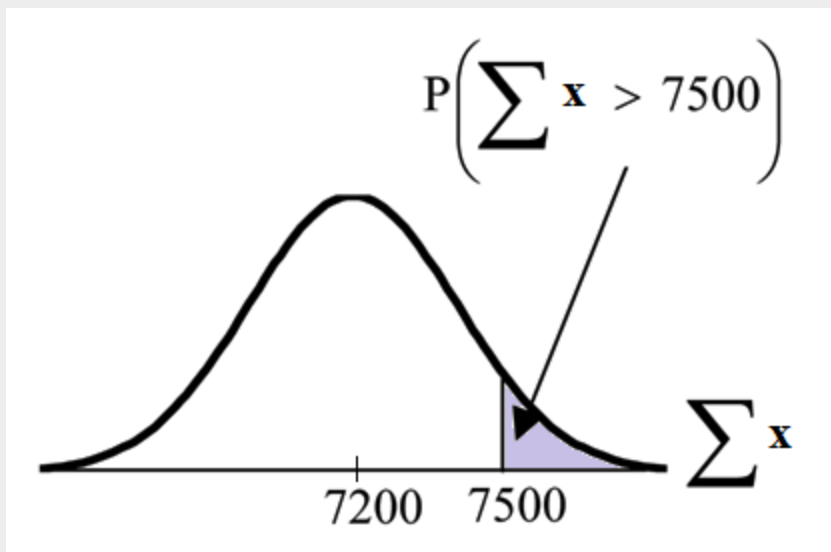
Let X = one value from the original unknown population. The probability question asks you to find a probability for **the sum (or total of) 80 values**.

ΣX = the sum or total of 80 values. Since $\mu_X = 90$, $\sigma_X = 15$, and $n = 80$, then

$$\Sigma X \sim N(80 \cdot 90, \sqrt{80} \cdot 15)$$

- mean of the sums = $n \cdot \mu_X = (80)(90) = 7200$
- standard deviation of the sums = $\sqrt{n} \cdot \sigma_X = \sqrt{80} \cdot 15$
- sum of 80 values = $\Sigma x = 7500$
- **a**Find $P(\Sigma x > 7500)$

$$P(\Sigma x > 7500) = 0.0127$$



normalcdf(lower value, upper value, mean of sums, **stdev** of sums)

The parameter list is abbreviated (lower, upper, $n \cdot \mu_X$, $\sqrt{n} \cdot \sigma_X$)

$$\text{normalcdf}(7500, 1E99, 80 \cdot 90, \sqrt{80} \cdot 15) = 0.0127$$

Reminder: $1E99 = 10^{99}$. Press the **EE** key for E.

- **bFind** Σx where $z = 1.5$:

$$\Sigma x = n \cdot \mu_X + z \cdot \sqrt{n} \cdot \sigma_X = (80)(90) + (1.5)(\sqrt{80})(15) = 7401.2$$

Glossary

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ . We are sampling with size n and we are interested in two new RVs - the sample mean, \bar{X} , and the sample sum, ΣX . If the size n of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(n\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Normal Distribution

A continuous random variable (RV) with pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \text{ where } \mu \text{ is the mean of the distribution and}$$

σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Using the Central Limit Theorem

Central Limit Theorem: Using the Central Limit Theorem is part of the collection col10555 written by Barbara Illowsky and Susan Dean. It covers how and when to use the Central Limit Theorem and has contributions from Roberta Bloom.

It is important for you to understand when to use the [CLT](#). If you are being asked to find the probability of the mean, use the CLT for the mean. If you are being asked to find the probability of a sum or total, use the CLT for sums. This also applies to percentiles for means and sums.

Note: If you are being asked to find the probability of an **individual** value, do **not** use the CLT. **Use the distribution of its random variable.**

Examples of the Central Limit Theorem

Law of Large Numbers

The [Law of Large Numbers](#) says that if you take samples of larger and larger size from any population, then the mean \bar{x} of the sample tends to get closer and closer to μ . From the Central Limit Theorem, we know that as n gets larger and larger, the sample means follow a normal distribution. The larger n gets, the smaller the standard deviation gets. (Remember that the standard deviation for \bar{X} is $\frac{\sigma}{\sqrt{n}}$.) This means that the sample mean \bar{x} must be close to the population mean μ . We can say that μ is the value that the sample means approach as n gets larger. The Central Limit Theorem illustrates the Law of Large Numbers.

Central Limit Theorem for the Mean and Sum Examples

Example:

A study involving stress is done on a college campus among the students. **The stress scores follow a uniform distribution** with the lowest stress score equal to 1 and the highest equal to 5. Using a sample of 75 students, find:

1. The probability that the **mean stress score** for the 75 students is less than 2.
2. The 90th percentile for the **mean stress score** for the 75 students.
3. The probability that the **total of the 75 stress scores** is less than 200.
4. The 90th percentile for the **total stress score** for the 75 students.

Let X = one stress score.

Problems 1. and 2. ask you to find a probability or a percentile for a **mean**. Problems 3 and 4 ask you to find a probability or a percentile for a **total or sum**. The sample size, n , is equal to 75.

Since the individual stress scores follow a uniform distribution, $X \sim U(1, 5)$ where $a = 1$ and $b = 5$ (See [Continuous Random Variables](#) for the uniform).

$$\mu_X = \frac{a+b}{2} = \frac{1+5}{2} = 3$$

$$\sigma_X = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(5-1)^2}{12}} = 1.15$$

For problems 1. and 2., let \bar{X} = the mean stress score for the 75 students. Then,

$$\bar{X} \sim N\left(3, \frac{1.15}{\sqrt{75}}\right) \quad \text{where } n = 75.$$

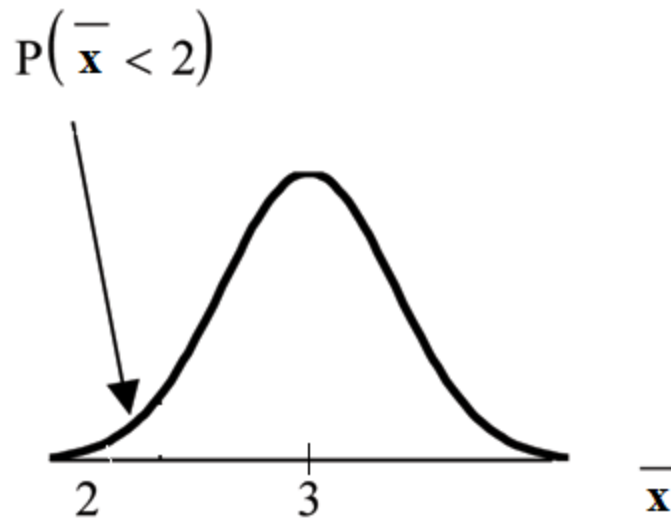
Exercise:

Problem: Find $P(\bar{x} < 2)$. Draw the graph.

Solution:

$$P(\bar{x} < 2) = 0$$

The probability that the mean stress score is less than 2 is about 0.



$$\text{normalcdf} \left(1, 2, 3, \frac{1.15}{\sqrt{75}} \right) = 0$$

Note: The smallest stress score is 1. Therefore, the smallest mean for 75 stress scores is 1.

Exercise:

Problem:

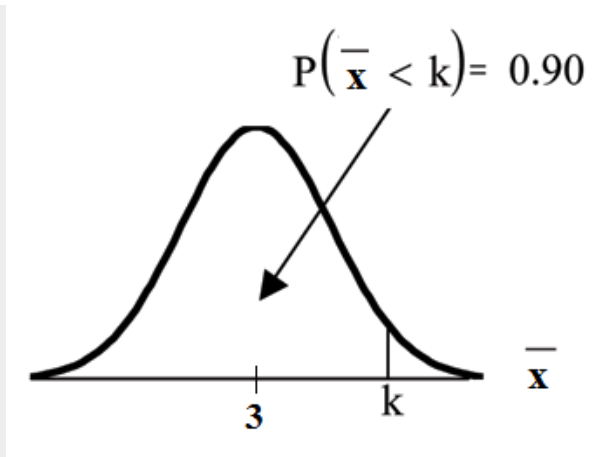
Find the 90th percentile for the mean of 75 stress scores. Draw a graph.

Solution:

Let k = the 90th percentile.

Find k where $P(x < k) = 0.90$.

$$k = 3.2$$



The 90th percentile for the mean of 75 scores is about 3.2. This tells us that 90% of all the means of 75 stress scores are at most 3.2 and 10% are at least 3.2.

$$\text{invNorm} \left(.90, 3, \frac{1.15}{\sqrt{75}} \right) = 3.2$$

For problems c and d, let ΣX = the sum of the 75 stress scores. Then, $\Sigma X \sim N \left[(75) \cdot (3), \sqrt{75} \cdot 1.15 \right]$

Exercise:

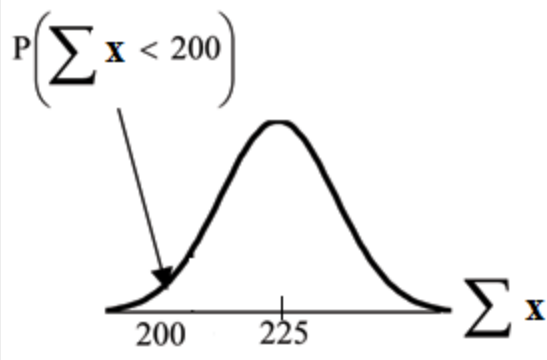
Problem: Find $P(\Sigma x < 200)$. Draw the graph.

Solution:

The mean of the sum of 75 stress scores is $75 \cdot 3 = 225$

The standard deviation of the sum of 75 stress scores is $\sqrt{75} \cdot 1.15 = 9.96$

$$P(\Sigma x < 200) = 0$$



The probability that the total of 75 scores is less than 200 is about 0.

$$\text{normalcdf} \left(75, 200, 75 \cdot 3, \sqrt{75} \cdot 1.15 \right) = 0.$$

Note: The smallest total of 75 stress scores is 75 since the smallest single score is 1.

Exercise:

Problem:

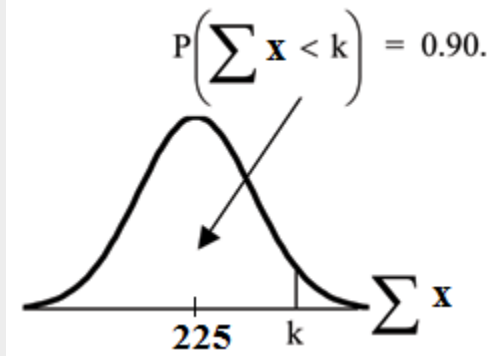
Find the 90th percentile for the total of 75 stress scores. Draw a graph.

Solution:

Let k = the 90th percentile.

Find k where $P(\Sigma x < k) = 0.90$.

$$k = 237.8$$



The 90th percentile for the sum of 75 scores is about 237.8. This tells us that 90% of all the sums of 75 scores are no more than 237.8 and 10% are no less than 237.8.

$$\text{invNorm} \left(.90, 75 \cdot 3, \sqrt{75} \cdot 1.15 \right) = 237.8$$

Example:

Suppose that a market research analyst for a cell phone company conducts a study of their customers who exceed the time allowance included on their basic cell phone contract; the analyst finds that for those people who exceed the time included in their basic contract, the **excess time used** follows an [exponential distribution](#) with a mean of 22 minutes.

Consider a random sample of 80 customers who exceed the time allowance included in their basic cell phone contract.

Let X = the excess time used by one INDIVIDUAL cell phone customer who exceeds his contracted time allowance.

$X \sim \text{Exp}\left(\frac{1}{22}\right)$ From Chapter 5, we know that $\mu = 22$ and $\sigma = 22$.

Let \bar{X} = the mean excess time used by a sample of $n = 80$ customers who exceed their contracted time allowance.

$\bar{X} \sim N\left(22, \frac{22}{\sqrt{80}}\right)$ by the CLT for Sample Means

Exercise:

Problem:

Using the CLT to find Probability:

- **a** Find the probability that the mean excess time used by the 80 customers in the sample is longer than 20 minutes. This is asking us to find $P(\bar{x} > 20)$ Draw the graph.
- **b** Suppose that one customer who exceeds the time limit for his cell phone contract is randomly selected. Find the probability that this individual customer's excess time is longer than 20 minutes. This is asking us to find $P(x > 20)$
- **c** Explain why the probabilities in (a) and (b) are different.

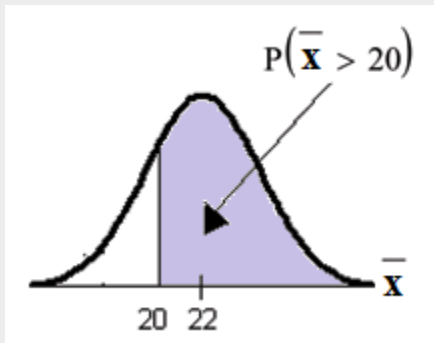
Solution:

Part a.

Find: $P(\bar{x} > 20)$

$$P(\bar{x} > 20) = 0.7919 \text{ using } \text{normalcdf} \left(20, 1E99, 22, \frac{22}{\sqrt{80}} \right)$$

The probability is 0.7919 that the mean excess time used is more than 20 minutes, for a sample of 80 customers who exceed their contracted time allowance.



Note: $1E99 = 10^{99}$ and $-1E99 = -10^{99}$. Press the

EE

key for E. Or just use 10^{99} instead of 1E99.

Part b.

Find $P(x > 20)$. Remember to use the exponential distribution for an **individual**: $X \sim \text{Exp}(1/22)$.

$$P(X > 20) = e^{-(1/22)*20} \text{ or } e^{(-.04545*20)} = 0.4029$$

Part c. Explain why the probabilities in (a) and (b) are different.

- $P(x > 20) = 0.4029$ but $P(\bar{x} > 20) = 0.7919$
- The probabilities are not equal because we use different distributions to calculate the probability for individuals and for means.
- When asked to find the probability of an individual value, use the stated distribution of its random variable; do not use the CLT. Use the CLT with the normal distribution when you are being asked to find the probability for an mean.

Exercise:**Problem:****Using the CLT to find Percentiles:**

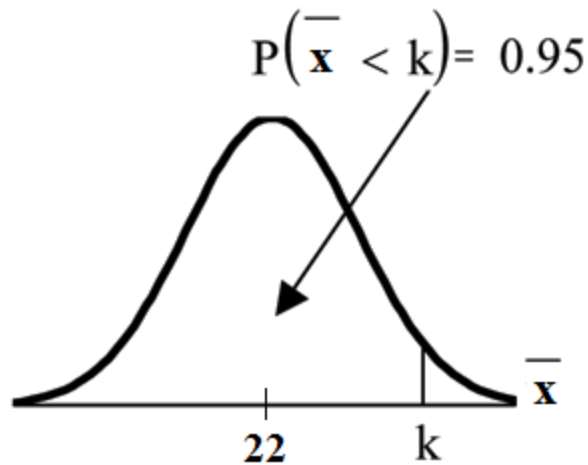
Find the 95th percentile for the **sample mean excess time** for samples of 80 customers who exceed their basic contract time allowances.

Draw a graph.

Solution:

Let k = the 95th percentile. Find k where $P(\bar{x} < k) = 0.95$

$$k = 26.0 \text{ using } \text{invNorm}\left(.95, 22, \frac{22}{\sqrt{80}}\right) = 26.0$$



The 95th percentile for the **sample mean excess time used** is about 26.0 minutes for random samples of 80 customers who exceed their contractual allowed time.

95% of such samples would have means under 26 minutes; only 5% of such samples would have means above 26 minutes.

Note:(HISTORICAL): Normal Approximation to the Binomial

Historically, being able to compute binomial probabilities was one of the most important applications of the Central Limit Theorem. Binomial probabilities were displayed in a table in a book with a small value for n (say, 20). To calculate the probabilities with large values of n , you had to use the binomial formula which could be very complicated. Using the [Normal Approximation to the Binomial](#) simplified the process. To compute the Normal Approximation to the Binomial, take a simple random sample from a population. You must meet the conditions for a **binomial distribution**:

- there are a certain number n of independent trials
- the outcomes of any trial are success or failure

- each trial has the same probability of a success p

Recall that if X is the binomial random variable, then $X \sim B(n, p)$. The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$; the approximation is better if they are both greater than or equal to 10). Then the binomial can be approximated by the normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$. Remember that $q = 1 - p$. In order to get the best approximation, add 0.5 to x or subtract 0.5 from x (use $x + 0.5$ or $x - 0.5$). The number 0.5 is called the continuity correction factor.

Example:

Suppose in a local Kindergarten through 12th grade (K - 12) school district, 53 percent of the population favor a charter school for grades K - 5. A simple random sample of 300 is surveyed.

1. Find the probability that **at least 150** favor a charter school.
2. Find the probability that **at most 160** favor a charter school.
3. Find the probability that **more than 155** favor a charter school.
4. Find the probability that **less than 147** favor a charter school.
5. Find the probability that **exactly 175** favor a charter school.

Let X = the number that favor a charter school for grades K - 5. $X \sim B(n, p)$ where $n = 300$ and $p = 0.53$. Since $np > 5$ and $nq > 5$, use the normal approximation to the binomial. The formulas for the mean and standard deviation are $\mu = np$ and $\sigma = \sqrt{npq}$. The mean is 159 and the standard deviation is 8.6447. The random variable for the normal distribution is Y . $Y \sim N(159, 8.6447)$. See **The Normal Distribution** for help with calculator instructions.

For Problem 1., you **include 150** so $P(x \geq 150)$ has normal approximation $P(Y \geq 149.5) = 0.8641$.

normalcdf (149.5, 10^99, 159, 8.6447) = 0.8641.

For Problem 2., you **include 160** so $P(x \leq 160)$ has normal approximation $P(Y \leq 160.5) = 0.5689$.

normalcdf (0, 160.5, 159, 8.6447) = 0.5689

For Problem 3., you **exclude 155** so $P(x > 155)$ has normal approximation $P(y > 155.5)=0.6572$.

normalcdf (155.5, 10⁹⁹, 159, 8.6447) = 0.6572

For Problem 4., you **exclude 147** so $P(x < 147)$ has normal approximation $P(Y < 146.5)=0.0741$.

normalcdf (0, 146.5, 159, 8.6447) = 0.0741

For Problem 5., $P(x=175)$ has normal approximation $P(174.5 < y < 175.5)=0.0083$.

normalcdf (174.5, 175.5, 159, 8.6447) = 0.0083

Because of calculators and computer software that easily let you calculate binomial probabilities for large values of n , it is not necessary to use the the Normal Approximation to the Binomial provided you have access to these technology tools. Most school labs have Microsoft Excel, an example of computer software that calculates binomial probabilities. Many students have access to the TI-83 or 84 series calculators and they easily calculate probabilities for the binomial. In an Internet browser, if you type in "binomial probability distribution calculation," you can find at least one online calculator for the binomial.

For **Example 3**, the probabilities are calculated using the binomial ($n=300$ and $p=0.53$) below. Compare the binomial and normal distribution answers. See **Discrete Random Variables** for help with calculator instructions for the binomial.

$P(x \geq 150)$: **1 - binomialcdf** (300, 0.53, 149)=0.8641

$P(x \leq 160)$: **binomialcdf** (300, 0.53, 160)=0.5684

$P(x > 155)$: **1 - binomialcdf** (300, 0.53, 155)=0.6576

$P(x < 147)$: **binomialcdf** (300, 0.53, 146)=0.0742

$P(x=175)$: (You use the binomial pdf.) **binomialpdf** (175, 0.53, 146)=0.0083

****Contributions made to Example 2 by Roberta Bloom**

Glossary

Average

A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ . We are sampling with size n and we are interested in two new RVs - the sample mean, \bar{X} , and the sample sum, ΣX . If the size n of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(n\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Exponential Distribution

A continuous random variable (RV) that appears when we are interested in the intervals of time between some random events, for example, the length of time between emergency arrivals at a hospital. Notation: $X \sim \text{Exp}(m)$. The mean is $\mu = \frac{1}{m}$ and the standard deviation is $\sigma = \frac{1}{m}$. The probability density function is $f(x) = me^{-mx}$, $x \geq 0$ and the cumulative distribution function is $P(X \leq x) = 1 - e^{-mx}$.

Mean

A number that measures the central tendency. A common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by \bar{x}) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Uniform Distribution

A continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$. Often referred as the **Rectangular distribution** because the graph of the pdf has the form of a rectangle. Notation: $X \sim U(a, b)$. The mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function is $f(x) = \frac{1}{b-a}$ for $a < x < b$ or $a \leq x \leq b$. The cumulative distribution is $P(X \leq x) = \frac{x-a}{b-a}$.

Summary of Formulas

Formula

Central Limit Theorem for Sample Means

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right) \quad \text{The Mean } (\bar{X}): \mu_X$$

Formula

Central Limit Theorem for Sample Means Z-Score and Standard Error of the Mean

$$z = \frac{\bar{x} - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)} \quad \text{Standard Error of the Mean (Standard Deviation } (\bar{X}): \frac{\sigma_X}{\sqrt{n}}$$

Formula

Central Limit Theorem for Sums

$$\Sigma X \sim N\left[(n) \cdot \mu_X, \sqrt{n} \cdot \sigma_X\right] \quad \text{Mean for Sums } (\Sigma X): n \cdot \mu_X$$

Formula

Central Limit Theorem for Sums Z-Score and Standard Deviation for Sums

$$z = \frac{\Sigma X - n \cdot \mu_X}{\sqrt{n} \cdot \sigma_X} \quad \text{Standard Deviation for Sums } (\Sigma X): \sqrt{n} \cdot \sigma_X$$

Practice: The Central Limit Theorem

Student Learning Outcomes

- The student will calculate probabilities using the Central Limit Theorem.

Given

Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately 4 hours each to do with a population standard deviation of 1.2 hours. Let X be the random variable representing the time it takes her to complete one review. Assume X is normally distributed. Let \bar{X} be the random variable representing the mean time to complete the 16 reviews. Let ΣX be the total time it takes Yoonie to complete all of the month's reviews. Assume that the 16 reviews represent a random set of reviews.

Distribution

Complete the distributions.

1. $X \sim$
2. $\bar{X} \sim$
3. $\Sigma X \sim$

Graphing Probability

For each problem below:

- **a** Sketch the graph. Label and scale the horizontal axis. Shade the region corresponding to the probability.
- **b** Calculate the value.

Exercise:

Problem:

Find the probability that **one** review will take Yoonie from 3.5 to 4.25 hours.

- a



- b $P(\text{_____} < x < \text{_____}) = \text{_____}$

Solution:

- b 3.5, 4.25, 0.2441

Exercise:**Problem:**

Find the probability that the **mean** of a month's reviews will take Yoonie from 3.5 to 4.25 hrs.

- a



- **b** $P(\text{_____}) = \text{_____}$

Solution:

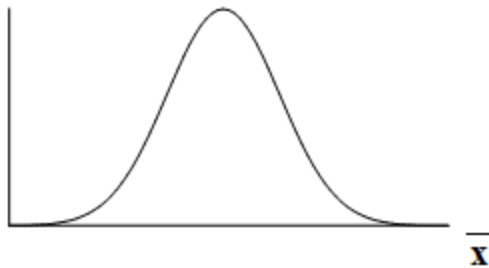
- **b** 0.7499

Exercise:

Problem:

Find the 95th percentile for the **mean** time to complete one month's reviews.

- **a**



- **b** The 95th Percentile =

Solution:

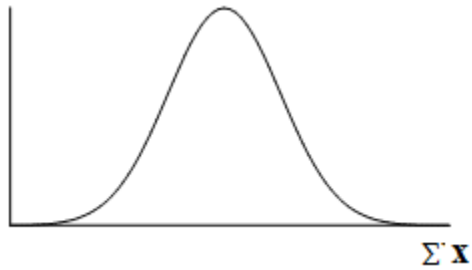
- **b** 4.49 hours

Exercise:

Problem:

Find the probability that the **sum** of the month's reviews takes Yoonie from 60 to 65 hours.

- **a**



- **b** The Probability=

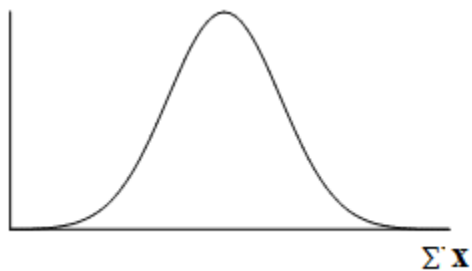
Solution:

- **b** 0.3802

Exercise:

Problem: Find the 95th percentile for the **sum** of the month's reviews.

- **a**



- **b** The 95th percentile=

Solution:

- **b** 71.90

Discussion Question

Exercise:

Problem: What causes the probabilities in [\[link\]](#) and [\[link\]](#) to differ?

Homework

The Central Limit Theorem: Homework is part of the collection col10555 written by Barbara Illowsky and Susan Dean.

Exercise:

Problem:

$X \sim N(60,9)$. Suppose that you form random samples of 25 from this distribution. Let \bar{X} be the random variable of averages. Let ΣX be the random variable of sums. For **c** - **f**, sketch the graph, shade the region, label and scale the horizontal axis for \bar{X} , and find the probability.

- **a** Sketch the distributions of \bar{X} and X on the same graph.
- **b** $\bar{X} \sim$
- **c** $P(\bar{x} < 60) =$
- **d** Find the 30th percentile for the mean.
- **e** $P(56 < \bar{x} < 62) =$
- **f** $P(18 < x < 58) =$
- **g** $\Sigma x \sim$
- **h** Find the minimum value for the upper quartile for the sum.
- **i** $P(1400 < \Sigma x < 1550) =$

Solution:

- **b** $\bar{X} \sim N(60, \frac{9}{\sqrt{25}})$
- **c** 0.5000
- **d** 59.06
- **e** 0.8536
- **f** 0.1333
- **h** 1530.35
- **i** 0.8536

Exercise:

Problem:

Determine which of the following are true and which are false. Then, in complete sentences, justify your answers.

- **a** When the sample size is large, the mean of \bar{X} is approximately equal to the mean of X .
- **b** When the sample size is large, \bar{X} is approximately normally distributed.
- **c** When the sample size is large, the standard deviation of \bar{X} is approximately the same as the standard deviation of X .

Exercise:**Problem:**

The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of about 10. Suppose that 16 individuals are randomly chosen.

Let \bar{X} = average percent of fat calories.

- **a** $\bar{X} \sim \text{_____} (\text{_____} , \text{_____})$
- **b** For the group of 16, find the probability that the average percent of fat calories consumed is more than 5. Graph the situation and shade in the area to be determined.
- **c** Find the first quartile for the average percent of fat calories.

Solution:

- **a** $N(36, \frac{10}{\sqrt{16}})$
- **b** 1
- **c** 34.31

Exercise:

Problem:

Previously, De Anza statistics students estimated that the amount of change daytime statistics students carry is exponentially distributed with a mean of \$0.88. Suppose that we randomly pick 25 daytime statistics students.

- **a** In words, $X =$
- **b** $X \sim$
- **c** In words, $\bar{X} =$
- **d** $\bar{X} \sim$ _____ (_____ , _____)
- **e** Find the probability that an individual had between \$0.80 and \$1.00. Graph the situation and shade in the area to be determined.
- **f** Find the probability that the average of the 25 students was between \$0.80 and \$1.00. Graph the situation and shade in the area to be determined.
- **g** Explain the why there is a difference in (e) and (f).

Exercise:**Problem:**

Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

- **a** If \bar{X} = average distance in feet for 49 fly balls, then $\bar{X} \sim$ _____ (_____ , _____)
- **b** What is the probability that the 49 balls traveled an average of less than 240 feet? Sketch the graph. Scale the horizontal axis for \bar{X} . Shade the region corresponding to the probability. Find the probability.
- **c** Find the 80th percentile of the distribution of the average of 49 fly balls.

Solution:

- **a** $N(250, \frac{50}{\sqrt{49}})$
- **b** 0.0808
- **c** 256.01 feet

Exercise:

Problem:

Suppose that the weight of open boxes of cereal in a home with children is uniformly distributed from 2 to 6 pounds. We randomly survey 64 homes with children.

- **a** In words, $X =$
- **b** $X \sim$
- **c** $\mu_X =$
- **d** $\sigma_X =$
- **e** In words, $\Sigma X =$
- **f** $\Sigma X \sim$
- **g** Find the probability that the total weight of open boxes is less than 250 pounds.
- **h** Find the 35th percentile for the total weight of open boxes of cereal.

Exercise:

Problem:

Suppose that the duration of a particular type of criminal trial is known to have a mean of 21 days and a standard deviation of 7 days. We randomly sample 9 trials.

- **a** In words, $\Sigma X =$
- **b** $\Sigma X \sim$
- **c** Find the probability that the total length of the 9 trials is at least 225 days.
- **d** 90 percent of the total of 9 of these types of trials will last at least how long?

Solution:

- **a** The total length of time for 9 criminal trials
- **b** $N(189, 21)$
- **c** 0.0432
- **d** 162.09

Exercise:**Problem:**

According to the Internal Revenue Service, the average length of time for an individual to complete (record keep, learn, prepare, copy, assemble and send) IRS Form 1040 is 10.53 hours (without any attached schedules). The distribution is unknown. Let us assume that the standard deviation is 2 hours. Suppose we randomly sample 36 taxpayers.

- **a** In words, $\bar{X} =$
- **b** In words, $\bar{X} =$
- **c** $\bar{X} \sim$
- **d** Would you be surprised if the 36 taxpayers finished their Form 1040s in an average of more than 12 hours? Explain why or why not in complete sentences.
- **e** Would you be surprised if one taxpayer finished his Form 1040 in more than 12 hours? In a complete sentence, explain why.

Exercise:**Problem:**

Suppose that a category of world class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races.

Let $\bar{X} =$ the average of the 49 races.

- **a** $X \sim$
 - **b** Find the probability that the runner will average between 142 and 146 minutes in these 49 marathons.
 - **c** Find the 80th percentile for the average of these 49 marathons.
 - **d** Find the median of the average running times.
-

Solution:

- **a** $N(145, \frac{14}{\sqrt{49}})$
- **b** 0.6247
- **c** 146.68
- **d** 145 minutes

Exercise:

Problem:

The attention span of a two year-old is exponentially distributed with a mean of about 8 minutes. Suppose we randomly survey 60 two year-olds.

- **a** In words, $X =$
- **b** $X \sim$
- **c** In words, $X =$
- **d** $X \sim$
- **e** Before doing any calculations, which do you think will be higher? Explain why.
 - **i** the probability that an individual attention span is less than 10 minutes; or
 - **ii** the probability that the average attention span for the 60 children is less than 10 minutes? Why?
- **f** Calculate the probabilities in part (e).
- **g** Explain why the distribution for X is not exponential.

Exercise:

Problem:

Suppose that the length of research papers is uniformly distributed from 10 to 25 pages. We survey a class in which 55 research papers were turned in to a professor. The 55 research papers are considered a random collection of all papers. We are interested in the average length of the research papers.

- **a** In words, $X =$
- **b** $X \sim$
- **c** $\mu_X =$
- **d** $\sigma_X =$
- **e** In words, $\bar{X} =$
- **f** $\bar{X} \sim$
- **g** In words, $\Sigma X =$
- **h** $\Sigma X \sim$
- **i** Without doing any calculations, do you think that it's likely that the professor will need to read a total of more than 1050 pages? Why?
- **j** Calculate the probability that the professor will need to read a total of more than 1050 pages.
- **k** Why is it so unlikely that the average length of the papers will be less than 12 pages?

Solution:

- **b** $U(10,25)$
- **c** 17.5
- **d** $\sqrt{\frac{225}{12}} = 4.3301$
- **f** $N(17.5, 0.5839)$
- **h** $N(962.5, 32.11)$
- **j** 0.0032

Exercise:

Problem:

The length of songs in a collector's CD collection is uniformly distributed from 2 to 3.5 minutes. Suppose we randomly pick 5 CDs from the collection. There is a total of 43 songs on the 5 CDs.

- **a** In words, \bar{X} =
- **b** $\bar{X} \sim$
- **c** In words, $\Sigma \bar{X}$ =
- **d** $\Sigma \bar{X} \sim$
- **e** Find the first quartile for the average song length.
- **f** The IQR (interquartile range) for the average song length is from _____ to _____.

Exercise:**Problem:**

Salaries for teachers in a particular elementary school district are normally distributed with a mean of \$44,000 and a standard deviation of \$6500. We randomly survey 10 teachers from that district.

- **a** In words, \bar{X} =
- **b** In words, $\Sigma \bar{X}$ =
- **c** $\bar{X} \sim$
- **d** In words, $\Sigma \bar{X} \sim$
- **e** $\Sigma \bar{X} \sim$
- **f** Find the probability that the teachers earn a total of over \$400,000.
- **g** Find the 90th percentile for an individual teacher's salary.
- **h** Find the 90th percentile for the average teachers' salary.
- **i** If we surveyed 70 teachers instead of 10, graphically, how would that change the distribution for \bar{X} ?
- **j** If each of the 70 teachers received a \$3000 raise, graphically, how would that change the distribution for \bar{X} ?

Solution:

- **c** $N(44,000, \frac{6500}{\sqrt{10}})$
- **e** $N(440,000, (\sqrt{10})(6500))$
- **f** 0.9742
- **g** \$52,330
- **h** \$46,634

Exercise:**Problem:**

The distribution of income in some Third World countries is considered wedge shaped (many very poor people, very few middle income people, and few to many wealthy people). Suppose we pick a country with a wedge distribution. Let the average salary be \$2000 per year with a standard deviation of \$8000. We randomly survey 1000 residents of that country.

- **a** In words, $X =$
- **b** In words, $X =$
- **c** $X \sim$
- **d** How is it possible for the standard deviation to be greater than the average?
- **e** Why is it more likely that the average of the 1000 residents will be from \$2000 to \$2100 than from \$2100 to \$2200?

Exercise:**Problem:**

The average length of a maternity stay in a U.S. hospital is said to be 2.4 days with a standard deviation of 0.9 days. We randomly survey 80 women who recently bore children in a U.S. hospital.

- **a** In words, $X =$
- **b** In words, $X =$

- **c** $X \sim$
- **d** In words, $\Sigma X =$
- **e** $\Sigma X \sim$
- **f** Is it likely that an individual stayed more than 5 days in the hospital? Why or why not?
- **g** Is it likely that the average stay for the 80 women was more than 5 days? Why or why not?
- **h** Which is more likely:
 - **i** an individual stayed more than 5 days; or
 - **ii** the average stay of 80 women was more than 5 days?
- **i** If we were to sum up the women's stays, is it likely that, collectively they spent more than a year in the hospital? Why or why not?

Solution:

- **c** $N(2.4, \frac{0.9}{\sqrt{80}})$
- **e** $N(192, 8.05)$
- **h** Individual

Exercise:

Problem:

In 1940 the average size of a U.S. farm was 174 acres. Let's say that the standard deviation was 55 acres. Suppose we randomly survey 38 farmers from 1940. (Source: U.S. Dept. of Agriculture)

- **a** In words, $X =$
- **b** In words, $\Sigma X =$
- **c** $X \sim$
- **d** The IQR for X is from _____ acres to _____ acres.

Exercise:

Problem:

The stock closing prices of 35 U.S. semiconductor manufacturers are given below. (Source: **Wall Street Journal**)

8.625 30.25 27.625 46.75 32.875 18.25 5 0.125 2.9375 6.875 28.25
24.25 21 1.5 30.25 71 43.5 49.25 2.5625 31 16.5 9.5 18.5 18 9 10.5
16.625 1.25 18 12.875 7 12.875 2.875 60.25 29.25

- **a** In words, $\bar{X} =$
- **b**
 - **i** $x =$
 - **ii** $s_x =$
 - **iii** $n =$
- **c** Construct a histogram of the distribution of the averages. Start at $x = -0.0005$. Make bar widths of 10.
- **d** In words, describe the distribution of stock prices.
- **e** Randomly average 5 stock prices together. (Use a random number generator.) Continue averaging 5 pieces together until you have 10 averages. List those 10 averages.
- **f** Use the 10 averages from (e) to calculate:
 - **i** $x =$
 - **ii** $s_x =$
- **g** Construct a histogram of the distribution of the averages. Start at $x = -0.0005$. Make bar widths of 10.
- **h** Does this histogram look like the graph in (c)?
- **i** In 1 - 2 complete sentences, explain why the graphs either look the same or look different?
- **j** Based upon the theory of the Central Limit Theorem, $\bar{X} \sim$

Solution:

- **b** \$20.71; \$17.31; 35

- **d** Exponential distribution, $X \sim \text{Exp}(1/20.71)$
- **f** \$20.71; \$11.14
- **j** $N(20.71, \frac{17.31}{\sqrt{5}})$

Exercise:

Problem:

Use the [Initial Public Offering data](#) (see “Table of Contents”) to do this problem.

- **a** In words, $X =$
- **b**
 - **i** $\mu_X =$
 - **ii** $\sigma_X =$
 - **iii** $n =$
- **c** Construct a histogram of the distribution. Start at $x = -0.50$. Make bar widths of \$5.
- **d** In words, describe the distribution of stock prices.
- **e** Randomly average 5 stock prices together. (Use a random number generator.) Continue averaging 5 pieces together until you have 15 averages. List those 15 averages.
- **f** Use the 15 averages from (e) to calculate the following:
 - **i** $\bar{x} =$
 - **ii** $s_x =$
- **g** Construct a histogram of the distribution of the averages. Start at $x = -0.50$. Make bar widths of \$5.
- **h** Does this histogram look like the graph in (c)? Explain any differences.
- **i** In 1 - 2 complete sentences, explain why the graphs either look the same or look different?
- **j** Based upon the theory of the Central Limit Theorem, $X \sim$

Try these multiple choice questions (Exercises 19 - 23).

The next two questions refer to the following information: The time to wait for a particular rural bus is distributed uniformly from 0 to 75 minutes. 100 riders are randomly sampled to learn how long they waited.

Exercise:

Problem:

The 90th percentile sample average wait time (in minutes) for a sample of 100 riders is:

- **A** 315.0
 - **B** 40.3
 - **C** 38.5
 - **D** 65.2
-

Solution:

B

Exercise:

Problem:

Would you be surprised, based upon numerical calculations, if the sample average wait time (in minutes) for 100 riders was less than 30 minutes?

- **A** Yes
 - **B** No
 - **C** There is not enough information.
-

Solution:

A

Exercise:

Problem:

Which of the following is NOT TRUE about the distribution for averages?

- **A** The mean, median and mode are equal
- **B** The area under the curve is one
- **C** The curve never touches the x-axis
- **D** The curve is skewed to the right

Solution:

D

The next three questions refer to the following information: The cost of unleaded gasoline in the Bay Area once followed an unknown distribution with a mean of \$4.59 and a standard deviation of \$0.10. Sixteen gas stations from the Bay Area are randomly chosen. We are interested in the average cost of gasoline for the 16 gas stations.

Exercise:**Problem:**

The distribution to use for the average cost of gasoline for the 16 gas stations is

- **A** $X \sim N(4.59, 0.10)$
- **B** $X \sim N\left(4.59, \frac{0.10}{\sqrt{16}}\right)$
- **C** $X \sim N\left(4.59, \frac{0.10}{16}\right)$
- **D** $X \sim N\left(4.59, \frac{16}{0.10}\right)$

Solution:

B

Exercise:

Problem:

What is the probability that the average price for 16 gas stations is over \$4.69?

- **A** Almost zero
 - **B** 0.1587
 - **C** 0.0943
 - **D** Unknown
-

Solution:

A

Exercise:

Problem:

Find the probability that the average price for 30 gas stations is less than \$4.55.

- **A**0.6554
 - **B**0.3446
 - **C**0.0142
 - **D**0.9858
 - **E**0
-

Solution:

C

Exercise:

Problem:

For the Charter School Problem (Example 6) in **Central Limit Theorem: Using the Central Limit Theorem**, calculate the following using the normal approximation to the binomial.

- **A** Find the probability that less than 100 favor a charter school for grades K - 5.
- **B** Find the probability that 170 or more favor a charter school for grades K - 5.
- **C** Find the probability that no more than 140 favor a charter school for grades K - 5.
- **D** Find the probability that there are fewer than 130 that favor a charter school for grades K - 5.
- **E** Find the probability that exactly 150 favor a charter school for grades K - 5.

If you either have access to an appropriate calculator or computer software, try calculating these probabilities using the technology. Try also using the suggestion that is at the bottom of **Central Limit Theorem: Using the Central Limit Theorem** for finding a website that calculates binomial probabilities.

Solution:

- **C** 0.0162
- **E** 0.0268

Exercise:**Problem:**

Four friends, Janice, Barbara, Kathy and Roberta, decided to carpool together to get to school. Each day the driver would be chosen by randomly selecting one of the four names. They carpool to school for 96 days. Use the normal approximation to the binomial to calculate the following probabilities. Round the standard deviation to 4 decimal places.

- **A** Find the probability that Janice is the driver at most 20 days.
- **B** Find the probability that Roberta is the driver more than 16 days.
- **C** Find the probability that Barbara drives exactly 24 of those 96 days.

If you either have access to an appropriate calculator or computer software, try calculating these probabilities using the technology. Try also using the suggestion that is at the bottom of **Central Limit Theorem: Using the Central Limit Theorem** for finding a website that calculates binomial probabilities.

Solution:

- **A** 0.2047
- **B** 0.9615
- **C** 0.0938

****Exercise 24 contributed by Roberta Bloom**

Review

Central Limit Theorem: Review is part of the collection col10555 written by Barbara Illowsky and Susan Dean. The module consists of review exercises.

The next three questions refer to the following information: Richard's Furniture Company delivers furniture from 10 A.M. to 2 P.M. continuously and uniformly. We are interested in how long (in hours) past the 10 A.M. start time that individuals wait for their delivery.

Exercise:

Problem: $X \sim$

- A $U(0,4)$
- B $U(10,2)$
- C $\text{Exp}(2)$
- D $N(2,1)$

Solution:

A

Exercise:

Problem: The average wait time is:

- A 1 hour
- B 2 hour
- C 2.5 hour
- D 4 hour

Solution:

B

Exercise:

Problem:

Suppose that it is now past noon on a delivery day. The probability that a person must wait at least $1\frac{1}{2}$ **more** hours is:

- **A** $\frac{1}{4}$
 - **B** $\frac{1}{2}$
 - **C** $\frac{3}{4}$
 - **D** $\frac{3}{8}$
-

Solution:

A

Exercise:

Problem: Given: $X \sim \text{Exp}(\frac{1}{3})$.

- **a** Find $P(x > 1)$
 - **b** Calculate the minimum value for the upper quartile.
 - **c** Find $P(x = \frac{1}{3})$
-

Solution:

- **a** 0.7165
- **b** 4.16
- **c** 0

Exercise:**Problem:**

- 40% of full-time students took 4 years to graduate
- 30% of full-time students took 5 years to graduate
- 20% of full-time students took 6 years to graduate

- 10% of full-time students took 7 years to graduate

The expected time for full-time students to graduate is:

- **A** 4 years
- **B** 4.5 years
- **C** 5 years
- **D** 5.5 years

Solution:

C

Exercise:

Problem:

Which of the following distributions is described by the following example?

Many people can run a short distance of under 2 miles, but as the distance increases, fewer people can run that far.

- **A** Binomial
- **B** Uniform
- **C** Exponential
- **D** Normal

Solution:

C

Exercise:

Problem:

The length of time to brush one's teeth is generally thought to be exponentially distributed with a mean of $\frac{3}{4}$ minutes. Find the probability that a randomly selected person brushes his/her teeth less than $\frac{3}{4}$ minutes.

- A 0.5
- B $\frac{3}{4}$
- C 0.43
- D 0.63

Solution:

D

Exercise:**Problem:**

Which distribution accurately describes the following situation?

The chance that a teenage boy regularly gives his mother a kiss goodnight (and he should!!) is about 20%. Fourteen teenage boys are randomly surveyed.

X = the number of teenage boys that regularly give their mother a kiss goodnight

- A $B(14, 0.20)$
- B $P(2.8)$
- C $N(2.8, 2.24)$
- D $\text{Exp}(\frac{1}{0.20})$

Solution:

A

Exercise:

Problem:

Which distribution accurately describes the following situation?

A 2008 report on technology use states that approximately 20 percent of U.S. households have never sent an e-mail. (source: <http://www.webguild.org/2008/05/20-percent-of-americans-have-never-used-email.php>) Suppose that we select a random sample of fourteen U.S. households .

X =the number of households in a 2008 sample of 14 households that have never sent an email

- A $B(14, 0.20)$
- B $P(2.8)$
- C $N(2.8, 2.24)$
- D $\text{Exp}(\frac{1}{0.20})$

Solution:

A

****Exercise 9 contributed by Roberta Bloom**

Lab 1: Central Limit Theorem (Pocket Change)

Class Time:

Names:

Student Learning Outcomes:

- The student will demonstrate and compare properties of the Central Limit Theorem.

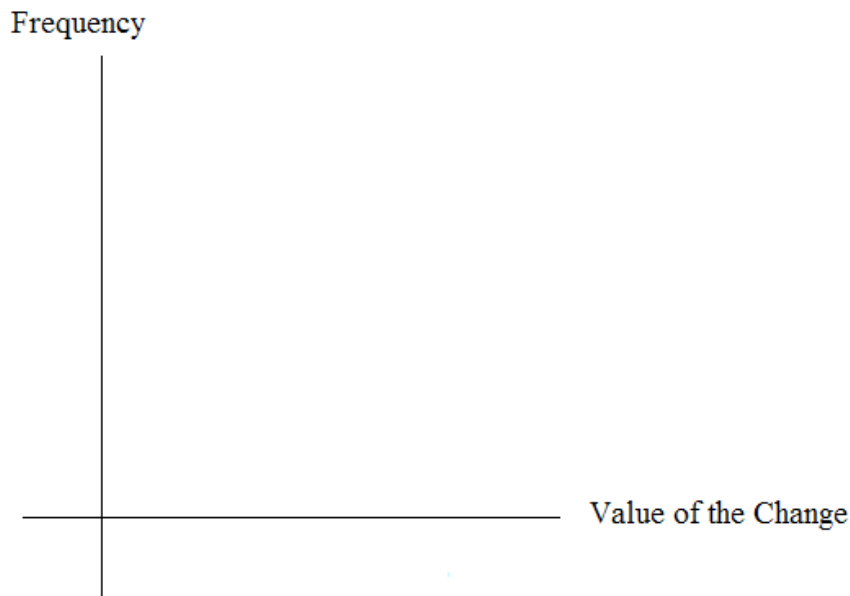
Note: This lab works best when sampling from several classes and combining data.

Collect the Data

1. Count the change in your pocket. (Do not include bills.)
2. Randomly survey 30 classmates. Record the values of the change.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

3. Construct a histogram. Make 5 - 6 intervals. Sketch the graph using a ruler and pencil.
Scale the axes.



4. Calculate the following ($n = 1$; surveying one person at a time):

- $\mathbf{a}\bar{x} =$
- $\mathbf{b}_s =$

5. Draw a smooth curve through the tops of the bars of the histogram. Use 1 – 2 complete sentences to describe the general shape of the curve.

Collecting Averages of Pairs

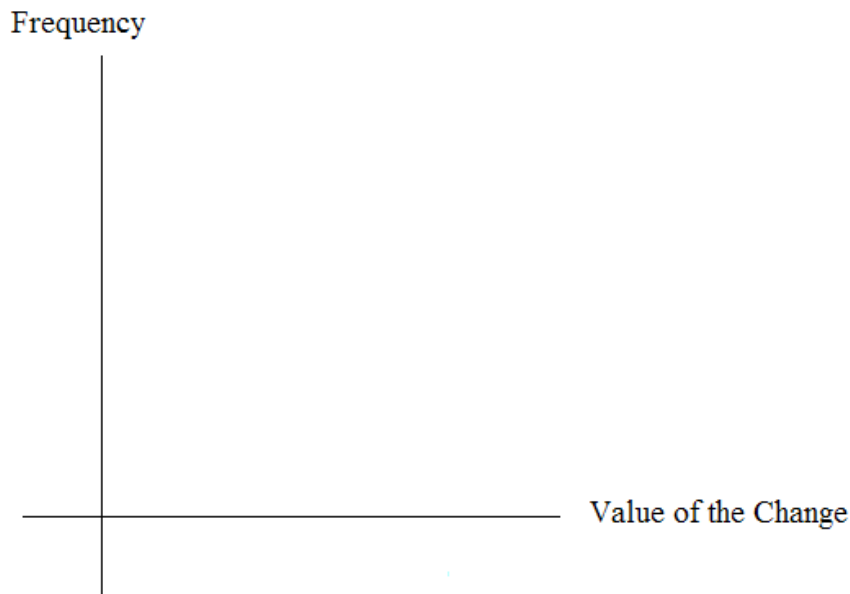
Repeat steps 1 - 5 (of the section above titled "Collect the Data") with one exception. Instead of recording the change of 30 classmates, record the average change of 30 pairs.

1. Randomly survey 30 **pairs** of classmates. Record the values of the average of their change.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

- Construct a histogram. Scale the axes using the same scaling you did for the section titled "Collecting the Data". Sketch the graph using a ruler and a pencil.



- Calculate the following ($n = 2$; surveying two people at a time):

- $\bar{a}x =$
- $\bar{b} s =$

- Draw a smooth curve through tops of the bars of the histogram. Use 1 – 2 complete sentences to describe the general shape of the curve.

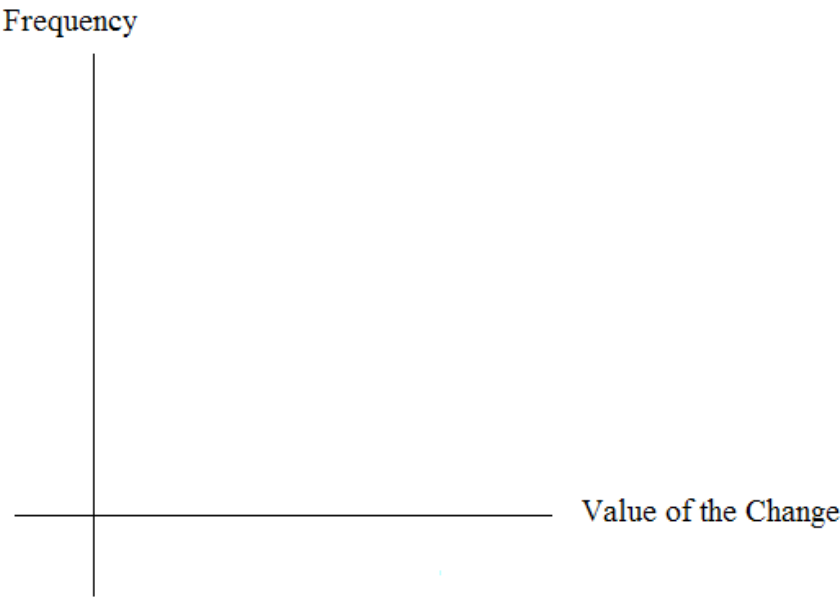
Collecting Averages of Groups of Five

Repeat steps 1 – 5 (of the section titled "Collect the Data") with one exception. Instead of recording the change of 30 classmates, record the average change of 30 groups of 5.

- Randomly survey 30 **groups of 5** classmates. Record the values of the average of their change.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

2. Construct a histogram. Scale the axes using the same scaling you did for the section titled "Collect the Data". Sketch the graph using a ruler and a pencil.



3. Calculate the following ($n = 5$; surveying five people at a time):

- $\mathbf{a} \bar{x} =$
- $\mathbf{b} s =$

4. Draw a smooth curve through tops of the bars of the histogram. Use 1 – 2 complete sentences to describe the general shape of the curve.

Discussion Questions

1. As n changed, why did the shape of the distribution of the data change? Use 1 – 2 complete sentences to explain what happened.
2. In the section titled "Collect the Data", what was the approximate distribution of the data? $X \sim$
3. In the section titled "Collecting Averages of Groups of Five", what was the approximate distribution of the averages? $\bar{X} \sim$
4. In 1 – 2 complete sentences, explain any differences in your answers to the previous two questions.

Lab 2: Central Limit Theorem (Cookie Recipes)

Class Time:

Names:

Student Learning Outcomes:

- The student will demonstrate and compare properties of the Central Limit Theorem.

Given:

X = length of time (in days) that a cookie recipe lasted at the Olmstead Homestead. (Assume that each of the different recipes makes the same quantity of cookies.)

Recipe #	X		Recipe #	X		Recipe #	X		Recipe #	X
1	1		16	2		31	3		46	2
2	5		17	2		32	4		47	2
3	2		18	4		33	5		48	11
4	5		19	6		34	6		49	5
5	6		20	1		35	6		50	5
6	1		21	6		36	1		51	4
7	2		22	5		37	1		52	6
8	6		23	2		38	2		53	5
9	5		24	5		39	1		54	1
10	2		25	1		40	6		55	1
11	5		26	6		41	1		56	2
12	1		27	4		42	6		57	4

Recipe #	X		Recipe #	X		Recipe #	X		Recipe #	X
13	1		28	1		43	2		58	3
14	3		29	6		44	6		59	6
15	2		30	2		45	2		60	5

Calculate the following:

- **a** $\mu_x =$
- **b** $\sigma_x =$

Collect the Data

Use a random number generator to randomly select 4 samples of size $n = 5$ from the given population. Record your samples below. Then, for each sample, calculate the mean to the nearest tenth. Record them in the spaces provided. Record the sample means for the rest of the class.

1. Complete the table:

	Sample 1	Sample 2	Sample 3	Sample 4	Sample means from other groups:
Means:	$\bar{x} =$	$\bar{x} =$	$\bar{x} =$	$\bar{x} =$	

2. Calculate the following:

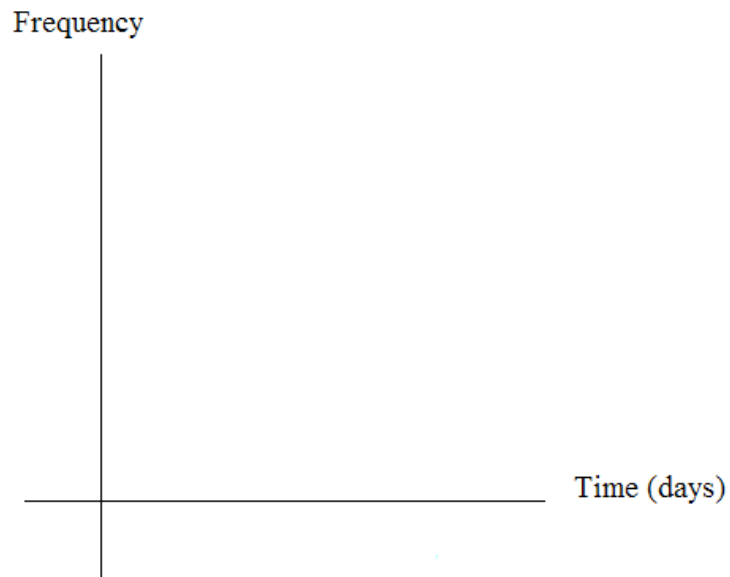
- **a** $\bar{x} =$
- **b** $s_{\bar{x}} =$

3. Again, use a random number generator to randomly select 4 samples from the population. This time, make the samples of size $n = 10$. Record the samples below. As before, for each

sample, calculate the mean to the nearest tenth. Record them in the spaces provided. Record the sample means for the rest of the class.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample means from other groups:
Means:	$\bar{x} =$	$\bar{x} =$	$\bar{x} =$	$\bar{x} =$	

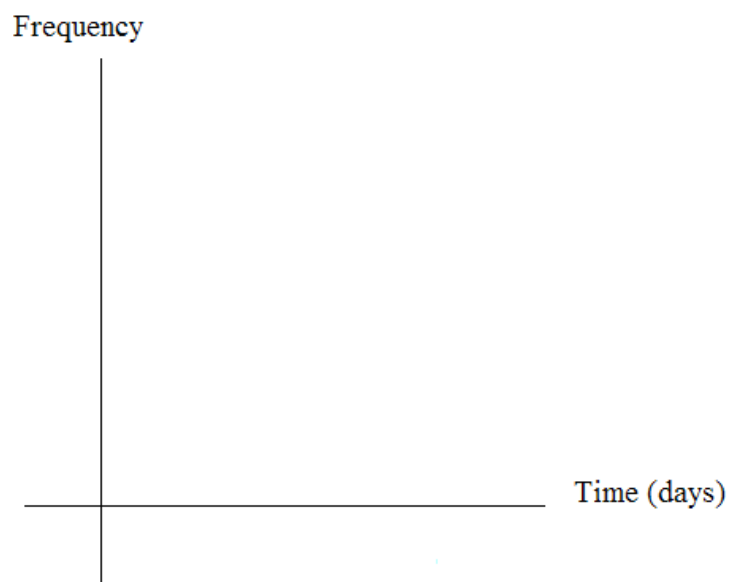
4. Calculate the following:
- $\mathbf{a}\bar{x} =$
 - $\mathbf{b}s_{\bar{x}} =$
5. For the original population, construct a histogram. Make intervals with bar width = 1 day. Sketch the graph using a ruler and pencil. Scale the axes.



6. Draw a smooth curve through the tops of the bars of the histogram. Use 1 – 2 complete sentences to describe the general shape of the curve.

Repeat the Procedure for $n=5$

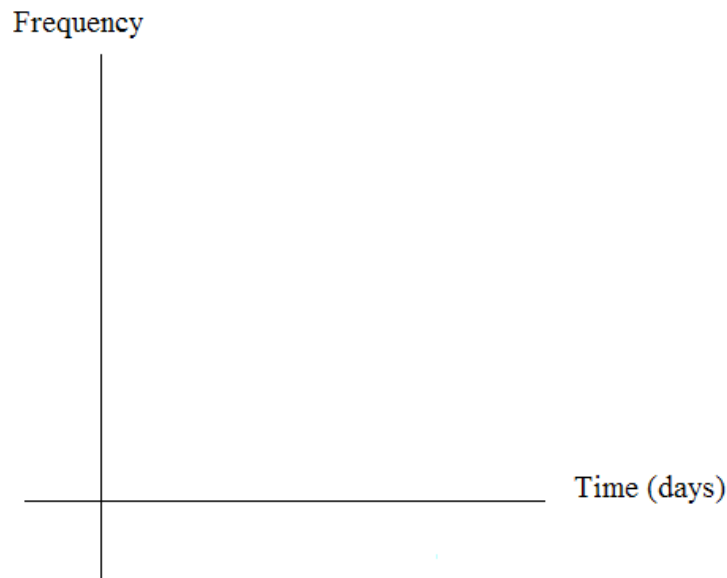
1. For the sample of $n = 5$ days averaged together, construct a histogram of the averages (your means together with the means of the other groups). Make intervals with bar widths $= \frac{1}{2}$ day. Sketch the graph using a ruler and pencil. Scale the axes.



2. Draw a smooth curve through the tops of the bars of the histogram. Use 1 – 2 complete sentences to describe the general shape of the curve.

Repeat the Procedure for $n=10$

1. For the sample of $n = 10$ days averaged together, construct a histogram of the averages (your means together with the means of the other groups). Make intervals with bar widths $= \frac{1}{2}$ day. Sketch the graph using a ruler and pencil. Scale the axes.



2. Draw a smooth curve through the tops of the bars of the histogram. Use 1 – 2 complete sentences to describe the general shape of the curve.

Discussion Questions

1. Compare the three histograms you have made, the one for the population and the two for the sample means. In three to five sentences, describe the similarities and differences.
2. State the theoretical (according to the CLT) distributions for the sample means.
 - **a** $n = 5$: $\bar{X} \sim$
 - **b** $n = 10$: $\bar{X} \sim$
3. Are the sample means for $n = 5$ and $n = 10$ “close” to the theoretical mean, μ_x ? Explain why or why not.
4. Which of the two distributions of sample means has the smaller standard deviation? Why?
5. As n changed, why did the shape of the distribution of the data change? Use 1 – 2 complete sentences to explain what happened.

Note: *This lab was designed and contributed by Carol Olmstead.*

Confidence Intervals

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Calculate and interpret confidence intervals for one population mean and one population proportion.
- Interpret the student-t probability distribution as the sample size changes.
- Discriminate between problems applying the normal and the student-t distributions.

Introduction

Suppose you are trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percent of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called [inferential statistics](#). **The sample data help us to make an estimate of a population [parameter](#).** We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct confidence intervals in which we believe the parameter lies.

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of compact discs (CD's) a consumer buys per month. If so, you could conduct a survey and calculate the sample mean, \bar{x} , and the sample standard deviation, s . You would use \bar{x} to estimate the population mean and s to estimate the population standard deviation. The sample mean, \bar{x} , is the **point estimate** for the population mean, μ . The sample standard deviation, s , is the point estimate for the population standard deviation, σ .

Each of \bar{x} and s is also called a statistic.

A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include an unknown population parameter.

Suppose for the CD example we do not know the population mean μ but we do know that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then by the Central Limit Theorem, the standard deviation for the sample mean is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1.$$

The **Empirical Rule**, which applies to bell-shaped distributions, says that in approximately 95% of the samples, the sample mean, \bar{x} , will be within two standard deviations of the population mean μ . For our CD example, two standard deviations is $(2)(0.1) = 0.2$. The sample mean \bar{x} is likely to be within 0.2 units of μ .

Because \bar{x} is within 0.2 units of μ , which is unknown, then μ is likely to be within 0.2 units of \bar{x} in 95% of the samples. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations $((2)(0.1))$ and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\bar{x} - 0.2$ and $\bar{x} + 0.2$ in 95% of all the samples.

For the CD example, suppose that a sample produced a sample mean $\bar{x} = 2$. Then the unknown population mean μ is between

$$\bar{x} - 0.2 = 2 - 0.2 = 1.8 \text{ and } \bar{x} + 0.2 = 2 + 0.2 = 2.2$$

We say that we are **95% confident** that the unknown population mean number of CDs is between 1.8 and 2.2. **The 95% confidence interval is (1.8, 2.2).**

The 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ or our sample produced an \bar{x} that is not within 0.2 units of the true mean μ . The second possibility happens for only 5% of all the samples (100% - 95%).

Remember that a confidence interval is created for an unknown population parameter like the population mean, μ . Confidence intervals for some parameters have the form

(point estimate - margin of error, point estimate + margin of error)

The margin of error depends on the confidence level or percentage of confidence.

When you read newspapers and journals, some reports will use the phrase "margin of error." Other reports will not use that phrase, but include a confidence interval as the point estimate + or - the margin of error. These are two ways of expressing the same concept.

Note: Although the text only covers symmetric confidence intervals, there are non-symmetric confidence intervals (for example, a confidence interval for the standard deviation).

Optional Collaborative Classroom Activity

Have your instructor record the number of meals each student in your class eats out in a week. Assume that the standard deviation is known to be 3 meals. Construct an approximate 95% confidence interval for the true mean number of meals students eat out each week.

1. Calculate the sample mean.
2. $\sigma = 3$ and $n =$ the number of students surveyed.
3. Construct the interval $\left(\bar{x} - 2 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 2 \cdot \frac{\sigma}{\sqrt{n}} \right)$

We say we are approximately 95% confident that the true average number of meals that students eat out in a week is between _____ and _____.

Glossary

Confidence Interval (CI)

An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Inferential Statistics

Also called statistical inference or inductive statistics. This facet of statistics deals with estimating a population parameter based on a sample statistic. For example, if 4 out of the 100 calculators sampled are defective we might infer that 4 percent of the production is defective.

Parameter

A numerical characteristic of the population.

Point Estimate

A single number computed from a sample and used to estimate a population parameter.

Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal

Confidence Intervals: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal is part of the collection col10555 written by Barbara Illowsky and Susan Dean with contributions from Roberta Bloom.

Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean μ , **where the population standard deviation is known**, we need x as an estimate for μ and we need the margin of error. Here, the margin of error is called the [error bound for a population mean](#) (abbreviated **EBM**). The sample mean x is the **point estimate** of the unknown population mean μ . **The confidence interval estimate will have the form:**

- (point estimate - error bound, point estimate + error bound) or, in symbols, $(x - \text{EBM}, x + \text{EBM})$

The margin of error depends on the [confidence level](#) (abbreviated **CL**). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, it is the choice of the person constructing the confidence interval to choose a confidence level of 90% or higher because that person wants to be reasonably certain of his or her conclusions.

There is another probability called alpha (α). α is related to the confidence level CL. α is the probability that the interval does not contain the unknown population parameter.

Mathematically, $\alpha + \text{CL} = 1$.

Example:

- Suppose we have collected data from a sample. We know the sample mean but we do not know the mean for the entire population.
- The sample mean is 7 and the error bound for the mean is 2.5.

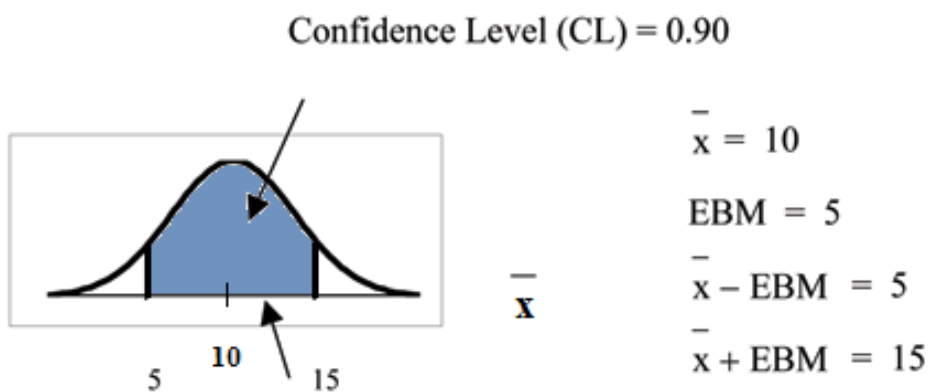
$x = 7$ and $EBM = 2.5$.

The confidence interval is $(7 - 2.5, 7 + 2.5)$; calculating the values gives $(4.5, 9.5)$.

If the confidence level (CL) is 95%, then we say that "We estimate with 95% confidence that the true value of the population mean is between 4.5 and 9.5."

A confidence interval for a population mean with a known standard deviation is based on the fact that the sample means follow an approximately normal distribution. Suppose that our sample has a mean of $x = 10$ and we have constructed the 90% confidence interval $(5, 15)$ where $EBM = 5$.

To get a 90% confidence interval, we must include the central 90% of the probability of the normal distribution. If we include the central 90%, we leave out a total of $\alpha = 10\%$ in both tails, or 5% in each tail, of the normal distribution.



μ is believed to be in the interval $(5, 15)$ with 90% confidence.

To capture the central 90%, we must go out 1.645 "standard deviations" on either side of the calculated sample mean. 1.645 is the z-score from a

Standard Normal probability distribution that puts an area of 0.90 in the center, an area of 0.05 in the far left tail, and an area of 0.05 in the far right tail.

It is important that the "standard deviation" used must be appropriate for the parameter we are estimating. So in this section, we need to use the standard deviation that applies to sample means, which is $\frac{\sigma}{\sqrt{n}}$. $\frac{\sigma}{\sqrt{n}}$ is commonly called the "standard error of the mean" in order to clearly distinguish the standard deviation for a mean from the population standard deviation σ .

In summary, as a result of the Central Limit Theorem:

- X is normally distributed, that is, $X \sim N\left(\mu_X, \frac{\sigma}{\sqrt{n}}\right)$.
- **When the population standard deviation σ is known, we use a Normal distribution to calculate the error bound.**

Calculating the Confidence Interval:

To construct a confidence interval estimate for an unknown population mean, we need data from a random sample. The steps to construct and interpret the confidence interval are:

- Calculate the sample mean \bar{x} from the sample data. Remember, in this section, we already know the population standard deviation σ .
- Find the Z-score that corresponds to the confidence level.
- Calculate the error bound EBM
- Construct the confidence interval
- Write a sentence that interprets the estimate in the context of the situation in the problem. (Explain what the confidence interval means, in the words of the problem.)

We will first examine each step in more detail, and then illustrate the process with some examples.

Finding z for the stated Confidence Level

When we know the population standard deviation σ , we use a standard normal distribution to calculate the error bound EBM and construct the confidence interval. We need to find the value of z that puts an area equal to

the confidence level (in decimal form) in the middle of the standard normal distribution $Z \sim N(0,1)$.

The confidence level, CL, is the area in the middle of the standard normal distribution. $CL = 1 - \alpha$. So α is the area that is split equally between the two tails. Each of the tails contains an area equal to $\frac{\alpha}{2}$.

The z-score that has an area to the right of $\frac{\alpha}{2}$ is denoted by $z_{\frac{\alpha}{2}}$

For example, when $CL = 0.95$ then $\alpha = 0.05$ and $\frac{\alpha}{2} = 0.025$; we write $z_{\frac{\alpha}{2}} = z_{0.025}$

The area to the right of $z_{0.025}$ is 0.025 and the area to the left of $z_{0.025}$ is $1 - 0.025 = 0.975$

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, using a calculator, computer or a Standard Normal probability table.

Using the TI83, TI83+ or TI84+ calculator: **invNorm**(0.975, 0, 1) = 1.96

CALCULATOR NOTE: Remember to use area to the LEFT of $z_{\frac{\alpha}{2}}$; in this chapter the last two inputs in the invNorm command are 0,1 because you are using a Standard Normal Distribution $Z \sim N(0,1)$

EBM: Error Bound

The error bound formula for an unknown population mean μ when the population standard deviation σ is known is

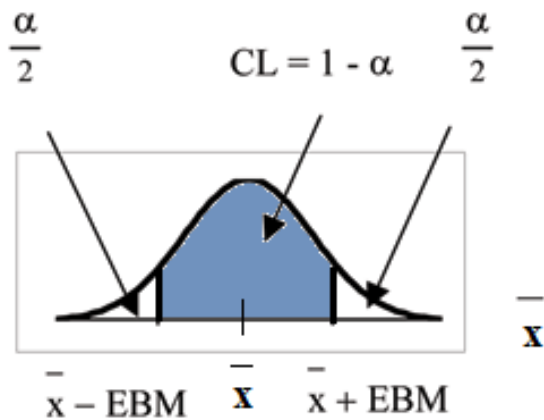
- $EBM = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

Constructing the Confidence Interval

- The confidence interval estimate has the format $(x - EBM, x + EBM)$.

The graph gives a picture of the entire situation.

$$CL + \frac{\alpha}{2} + \frac{\alpha}{2} = CL + \alpha = 1.$$



Writing the Interpretation

The interpretation should clearly state the confidence level (CL), explain what population parameter is being estimated (here, a **population mean**), and should state the confidence interval (both endpoints). "We estimate with ____% confidence that the true population mean (include context of the problem) is between ____ and ____ (include appropriate units)."

Example:

Suppose scores on exams in statistics are normally distributed with an unknown population mean and a population standard deviation of 3 points. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Exercise:

Problem:

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

Solution:

- You can use technology to directly calculate the confidence interval
- The first solution is shown step-by-step (Solution A).
- The second solution uses the TI-83, 83+ and 84+ calculators (Solution B).

Solution A

To find the confidence interval, you need the sample mean, \bar{x} , and the EBM.

- $\bar{x} = 68$
- $EBM = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}} \right)$
- $\sigma = 3$; $n = 36$; The confidence level is 90% (CL=0.90)

$$CL = 0.90 \text{ so } \alpha = 1 - CL = 1 - 0.90 = 0.10$$

$$\frac{\alpha}{2} = 0.05 \quad z_{\frac{\alpha}{2}} = z_{.05}$$

The area to the right of $z_{.05}$ is 0.05 and the area to the left of $z_{.05}$ is $1-0.05=0.95$

$$z_{\frac{\alpha}{2}} = z_{.05} = 1.645$$

using invNorm(0.95,0,1) on the TI-83,83+,84+ calculators. This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the Standard Normal distribution.

$$EBM = 1.645 \cdot \left(\frac{3}{\sqrt{36}} \right) = 0.8225$$

$$\bar{x} - EBM = 68 - 0.8225 = 67.1775$$

$$\bar{x} + EBM = 68 + 0.8225 = 68.8225$$

The 90% confidence interval is **(67.1775, 68.8225)**.

Solution B

Using a function of the TI-83, TI-83+ or TI-84 calculators:

Press **STAT** and arrow over to **TESTS**.

Arrow down to **7:ZInterval**.

Press **ENTER**.

Arrow to **Stats** and press **ENTER**.

Arrow down and enter 3 for σ , 68 for \bar{x} , 36 for n , and .90 for **C-level**.

Arrow down to **Calculate** and press **ENTER**.

The confidence interval is (to 3 decimal places) (67.178, 68.822).

Interpretation

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

Explanation of 90% Confidence Level

90% of all confidence intervals constructed in this way contain the true mean statistics exam score. For example, if we constructed 100 of these confidence intervals, we would expect 90 of them to contain the true population mean exam score.

Changing the Confidence Level or Sample Size

Example: Changing the Confidence Level

Exercise:

Problem:

Suppose we change the original problem by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Solution:

To find the confidence interval, you need the sample mean, \bar{x} , and the EBM.

- $x = 68$
- $EBM = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}} \right)$
- $\sigma = 3$; $n = 36$; The confidence level is 95% (CL=0.95)

$$CL = 0.95 \text{ so } \alpha = 1 - CL = 1 - 0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025 \quad z_{\frac{\alpha}{2}} = z_{.025}$$

The area to the right of $z_{.025}$ is 0.025 and the area to the left of $z_{.025}$ is $1-0.025=0.975$

$$z_{\frac{\alpha}{2}} = z_{.025} = 1.96$$

using invnorm(.975,0,1) on the TI-83,83+,84+ calculators. (This can also be found using appropriate commands on other calculators, using a computer, or using a probability table for the Standard Normal distribution.)

$$EBM = 1.96 \cdot \left(\frac{3}{\sqrt{36}} \right) = 0.98$$

$$x - EBM = 68 - 0.98 = 67.02$$

$$x + EBM = 68 + 0.98 = 68.98$$

Interpretation

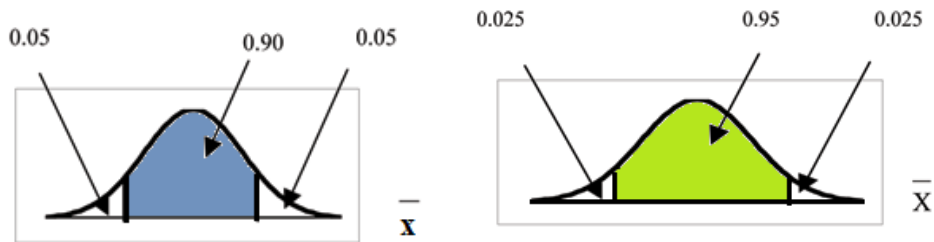
We estimate with 95 % confidence that the true population mean for all statistics exam scores is between 67.02 and 68.98.

Explanation of 95% Confidence Level

95% of all confidence intervals constructed in this way contain the true value of the population mean statistics exam score.

Comparing the results

The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider.



Summary: Effect of Changing the Confidence Level

- Increasing the confidence level increases the error bound, making the confidence interval wider.
- Decreasing the confidence level decreases the error bound, making the confidence interval narrower.

Example: Changing the Sample Size:

Suppose we change the original problem to see what happens to the error bound if the sample size is changed.

Exercise:

Problem:

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the error bound and the confidence interval if we increase the sample size and use $n=100$ instead of $n=36$? What happens if we decrease the sample size to $n=25$ instead of $n=36$?

- $x = 68$
- $EBM = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}} \right)$
- $\sigma = 3$; The confidence level is 90% (CL=0.90) ;
 $z_{\frac{\alpha}{2}} = z_{.05} = 1.645$

Solution:

If we **increase** the sample size n to 100, we **decrease** the error bound.

$$\text{When } n = 100 : \text{EBM} = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}} \right) = 1.645 \cdot \left(\frac{3}{\sqrt{100}} \right) = 0.4935$$

Solution:

If we **decrease** the sample size n to 25, we **increase** the error bound.

$$\text{When } n = 25 : \text{EBM} = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}} \right) = 1.645 \cdot \left(\frac{3}{\sqrt{25}} \right) = 0.987$$

Summary: Effect of Changing the Sample Size

- Increasing the sample size causes the error bound to decrease, making the confidence interval narrower.
- Decreasing the sample size causes the error bound to increase, making the confidence interval wider.

Working Backwards to Find the Error Bound or Sample Mean

Working Backwards to find the Error Bound or the Sample Mean

When we calculate a confidence interval, we find the sample mean and calculate the error bound and use them to calculate the confidence interval. But sometimes when we read statistical studies, the study may state the confidence interval only. If we know the confidence interval, we can work backwards to find both the error bound and the sample mean.

Finding the Error Bound

- From the upper value for the interval, subtract the sample mean
- OR, From the upper value for the interval, subtract the lower value. Then divide the difference by 2.

Finding the Sample Mean

- Subtract the error bound from the upper value of the confidence interval
- OR, Average the upper and lower endpoints of the confidence interval

Notice that there are two methods to perform each calculation. You can choose the method that is easier to use with the information you know.

Example:

Suppose we know that a confidence interval is **(67.18, 68.82)** and we want to find the error bound. We may know that the sample mean is 68. Or perhaps our source only gave the confidence interval and did not tell us the value of the sample mean.

Calculate the Error Bound:

- If we know that the sample mean is 68: $EBM = 68.82 - 68 = 0.82$
- If we don't know the sample mean: $EBM = \frac{(68.82 - 67.18)}{2} = 0.82$

Calculate the Sample Mean:

- If we know the error bound: $x = 68.82 - 0.82 = 68$
- If we don't know the error bound: $x = \frac{(67.18 + 68.82)}{2} = 68$

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population mean when the population standard deviation is known is $EBM = z_{\frac{\alpha}{2}} \cdot \left(\frac{\sigma}{\sqrt{n}} \right)$

The formula for sample size is $n = \frac{z^2 \sigma^2}{EBM^2}$, found by solving the error bound formula for n

In this formula, z is $z_{\frac{\alpha}{2}}$, corresponding to the desired confidence level. A researcher planning a study who wants a specified confidence level and error bound can use this formula to calculate the size of the sample needed for the study.

Example:

The population standard deviation for the age of Foothill College students is 15 years. If we want to be 95% confident that the sample mean age is within 2 years of the true population mean age of Foothill College students, how many randomly selected Foothill College students must be surveyed?

- From the problem, we know that $\sigma = 15$ and $EBM=2$
- $z = z_{.025} = 1.96$, because the confidence level is 95%.
- $n = \frac{z^2\sigma^2}{EBM^2} = \frac{1.96^2 15^2}{2^2} = 216.09$ using the sample size equation.
- Use $n = 217$: Always round the answer UP to the next higher integer to ensure that the sample size is large enough.

Therefore, 217 Foothill College students should be surveyed in order to be 95% confident that we are within 2 years of the true population mean age of Foothill College students.

**With contributions from Roberta Bloom

Glossary

Confidence Interval (CI)

An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.

- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Confidence Level (CL)

The percent expression for the probability that the confidence interval contains the true population parameter. For example, if the $CL = 90\%$, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Error Bound for a Population Mean (EBM)

The margin of error. Depends on the confidence level, sample size, and known or estimated population standard deviation.

Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student-T

Confidence Interval, Single Population Mean, Population Standard Deviation Unknown, Student-t is part of the collection col10555 written by Barbara Illowsky and Susan Dean with contributions from Roberta Bloom.

In practice, we rarely know the population [standard deviation](#). In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation s as an estimate for σ and proceeded as before to calculate a [confidence interval](#) with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Gossett (1876-1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with s did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the [Student's-t distribution](#). The name comes from the fact that Gosset wrote under the pen name "Student."

Up until the mid 1970s, some statisticians used the [normal distribution](#) approximation for large sample sizes and only used the Student's-t distribution for sample sizes of at most 30. With the common use of graphing calculators and computers, the practice is to use the Student's-t distribution whenever s is used as an estimate for σ .

If you draw a simple random sample of size n from a population that has approximately a normal distribution with mean μ and unknown population standard deviation σ and calculate the t-score $t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$, then the t-scores

follow a **Student's-t distribution with $n - 1$ degrees of freedom**. The t-score has the same interpretation as the [z-score](#). It measures how far \bar{x} is from its mean μ . For each sample size n , there is a different Student's-t distribution.

The **degrees of freedom**, $n - 1$, come from the calculation of the sample standard deviation s . In Chapter 2, we used n deviations ($x - \bar{x}$ values) to calculate s . Because the sum of the deviations is 0, we can find the last deviation once we know the other $n - 1$ deviations. The other $n - 1$ deviations can change or vary freely. **We call the number $n - 1$ the degrees of freedom (df).**

Properties of the Student's-t Distribution

- The graph for the Student's-t distribution is similar to the Standard Normal curve.
- The mean for the Student's-t distribution is 0 and the distribution is symmetric about 0.
- The Student's-t distribution has more probability in its tails than the Standard Normal distribution because the spread of the t distribution is greater than the spread of the Standard Normal. So the graph of the Student's-t distribution will be thicker in the tails and shorter in the center than the graph of the Standard Normal distribution.
- The exact shape of the Student's-t distribution depends on the "degrees of freedom". As the degrees of freedom increases, the graph Student's-t distribution becomes more like the graph of the Standard Normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ . The size of the underlying population is generally not relevant unless it is very small. If it is bell shaped (normal) then the assumption is met and doesn't need discussion. Random sampling is assumed but it is a completely separate assumption from normality.

Calculators and computers can easily calculate any Student's-t probabilities. The TI-83,83+,84+ have a tcdf function to find the probability for given values of t. The grammar for the tcdf command is tcdf(lower bound, upper bound, degrees of freedom). However for confidence intervals, we need to use **inverse** probability to find the value of t when we know the probability.

For the TI-84+ you can use the invT command on the DISTRibution menu. The invT command works similarly to the invnorm. The invT command

requires two inputs: **invT(area to the left, degrees of freedom)** The output is the t-score that corresponds to the area we specified.

The TI-83 and 83+ do not have the invT command. (The TI-89 has an inverse T command.)

A probability table for the Student's-t distribution can also be used. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). (The TI-86 does not have an invT program or command, so if you are using that calculator, you need to use a probability table for the Student's-t distribution.) When using t-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails.

A Student's-t table (See the Table of Contents **15. Tables**) gives t-scores given the degrees of freedom and the right-tailed probability. The table is very limited. **Calculators and computers can easily calculate any Student's-t probabilities.**

The notation for the Student's-t distribution is (using T as the random variable) is

- $T \sim t_{df}$ where $df = n - 1$.
- For example, if we have a sample of size $n=20$ items, then we calculate the degrees of freedom as $df=n-1=20-1=19$ and we write the distribution as $T \sim t_{19}$

If the population standard deviation is not known, the [error bound for a population mean](#) is:

- $EBM = t_{\frac{\alpha}{2}} \cdot \left(\frac{s}{\sqrt{n}} \right)$
- $t_{\frac{\alpha}{2}}$ is the t-score with area to the right equal to $\frac{\alpha}{2}$
- use $df = n - 1$ degrees of freedom
- s = sample standard deviation

The format for the confidence interval is:

$$(\bar{x} - \text{EBM}, \bar{x} + \text{EBM}).$$

The TI-83, 83+ and 84 calculators have a function that calculates the confidence interval directly. To get to it,

Press **STAT**

Arrow over to **TESTS**.

Arrow down to **8: TInterval** and press **ENTER** (or just press **8**).

Example:

Exercise:

Problem:

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given below. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data.

The solution is shown step-by-step and by using the TI-83, 83+ and 84+ calculators.

8.6 9.4 7.9 6.8 8.3 7.3 9.2 9.6 8.7 11.4 10.3 5.4 8.1 5.5 6.9

Solution:

- You can use technology to directly calculate the confidence interval.
- The first solution is step-by-step (Solution A).
- The second solution uses the Ti-83+ and Ti-84 calculators (Solution B).

Solution A

To find the confidence interval, you need the sample mean, \bar{x} , and the EBM.

$$\bar{x} = 8.2267 \quad s = 1.6722 \quad n = 15$$

$$df = 15 - 1 = 14$$

$$CL = 0.95 \quad \text{so} \quad \alpha = 1 - CL = 1 - 0.95 = 0.05$$

$$\frac{\alpha}{2} = 0.025 \quad t_{\frac{\alpha}{2}} = t_{.025}$$

The area to the right of $t_{.025}$ is 0.025 and the area to the left of $t_{.025}$ is $1 - 0.025 = 0.975$

$t_{\frac{\alpha}{2}} = t_{.025} = 2.14$ using $\text{invT}(.975, 14)$ on the TI-84+ calculator.

$$EBM = t_{\frac{\alpha}{2}} \cdot \left(\frac{s}{\sqrt{n}} \right)$$

$$EBM = 2.14 \cdot \left(\frac{1.6722}{\sqrt{15}} \right) = 0.924$$

$$\bar{x} - EBM = 8.2267 - 0.9240 = 7.3$$

$$\bar{x} + EBM = 8.2267 + 0.9240 = 9.15$$

The 95% confidence interval is **(7.30, 9.15)**.

We estimate with 95% confidence that the true population mean sensory rate is between 7.30 and 9.15.

Solution B

Using a function of the TI-83, TI-83+ or TI-84 calculators:

Press **STAT** and arrow over to **TESTS**.

Arrow down to **8:Interval** and press **ENTER** (or you can just press **8**). Arrow to **Data** and press **ENTER**.

Arrow down to **List** and enter the list name where you put the data.

Arrow down to **Freq** and enter 1.

Arrow down to **C-level** and enter .95

Arrow down to **Calculate** and press **ENTER**.

The 95% confidence interval is (7.3006, 9.1527)

Note:When calculating the error bound, a probability table for the Student's-t distribution can also be used to find the value of t. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row); the t-score is found where the row and column intersect in the table.

****With contributions from Roberta Bloom**

Glossary

Confidence Interval (CI)

An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Confidence Level (CL)

The percent expression for the probability that the confidence interval contains the true population parameter. For example, if the $CL = 90\%$, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Degrees of Freedom (df)

The number of objects in a sample that are free to vary.

Error Bound for a Population Mean (EBM)

The margin of error. Depends on the confidence level, sample size, and known or estimated population standard deviation.

Normal Distribution

A continuous random variable (RV) with pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \text{ where } \mu \text{ is the mean of the distribution and}$$

σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Student's-t Distribution

Investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data.

Confidence Interval for a Population Proportion

Confidence Interval for a Population Proportion is part of the collection col10555 written by Barbara Illowsky and Susan Dean with contributions from Roberta Bloom.

During an election year, we see articles in the newspaper that state [confidence intervals](#) in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within 3 percentage points. Often, election polls are calculated with 95% confidence. So, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43 : $(0.40 - 0.03, 0.40 + 0.03)$.

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval, the sample size, the [error bound](#), and the [confidence level](#) for a proportion is similar to that for the population mean. The formulas are different.

How do you know you are dealing with a proportion problem? First, the underlying [distribution is binomial](#). (There is no mention of a mean or average.) If X is a binomial random variable, then $X \sim B(n, p)$ where n = the number of trials and p = the probability of a success. To form a proportion, take X , the random variable for the number of successes and divide it by n , the number of trials (or the sample size). The random variable P' (read "P prime") is that proportion,

$$P' = \frac{X}{n}$$

(Sometimes the random variable is denoted as \hat{P} , read "P hat".)

When n is large and p is not close to 0 or 1, we can use the **normal distribution** to approximate the binomial.

$$X \sim N(n \cdot p, \sqrt{n \cdot p \cdot q})$$

If we divide the random variable by n , the mean by n , and the standard deviation by n , we get a normal distribution of proportions with P' , called the estimated proportion, as the random variable. (Recall that a proportion = the number of successes divided by n .)

$$\frac{X}{n} = P' \sim N\left(\frac{n \cdot p}{n}, \frac{\sqrt{n \cdot p \cdot q}}{n}\right)$$

Using algebra to simplify : $\frac{\sqrt{n \cdot p \cdot q}}{n} = \sqrt{\frac{p \cdot q}{n}}$

P' follows a normal distribution for proportions: $P' \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$

The confidence interval has the form $(p' - \text{EBP}, p' + \text{EBP})$.

$$p' = \frac{x}{n}$$

p' = the **estimated proportion** of successes (p' is a **point estimate** for p , the true proportion)

x = the **number** of successes.

n = the size of the sample

The error bound for a proportion is

$$\text{EBP} = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}} \quad \text{where } q' = 1 - p'$$

This formula is similar to the error bound formula for a mean, except that the "appropriate standard deviation" is different. For a mean, when the population standard deviation is known, the appropriate standard deviation

that we use is $\frac{\sigma}{\sqrt{n}}$. For a proportion, the appropriate standard deviation is $\sqrt{\frac{p \cdot q}{n}}$.

However, in the error bound formula, we use $\sqrt{\frac{p' \cdot q'}{n}}$ as the standard deviation, instead of $\sqrt{\frac{p \cdot q}{n}}$

However, in the error bound formula, the standard deviation is $\sqrt{\frac{p' \cdot q'}{n}}$.

In the error bound formula, the **sample proportions p' and q' are estimates of the unknown population proportions p and q** . The estimated proportions p' and q' are used because p and q are not known. p' and q' are calculated from the data. p' is the estimated proportion of successes. q' is the estimated proportion of failures.

The confidence interval can only be used if the number of successes np' and the number of failures nq' are both larger than 5.

Note:For the normal distribution of proportions, the z-score formula is as follows.

If $P' \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$ then the z-score formula is $z = \frac{p' - p}{\sqrt{\frac{p \cdot q}{n}}}$

Example:

Exercise:

Problem:

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have cell phones. 500 randomly selected adult residents in this city are surveyed to determine whether they have cell phones. Of the 500 people surveyed, 421 responded yes - they own cell phones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adults residents of this city who have cell phones.

Solution

- You can use technology to directly calculate the confidence interval.
- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

Solution:

Let X = the number of people in the sample who have cell phones. X is binomial. $X \sim B(500, \frac{421}{500})$.

To calculate the confidence interval, you must find p' , q' , and EBP.

$$n = 500 \quad x = \text{the number of successes} = 421$$

$$p' = \frac{x}{n} = \frac{421}{500} = 0.842$$

$p' = 0.842$ is the sample proportion; this is the point estimate of the population proportion.

$$q' = 1 - p' = 1 - 0.842 = 0.158$$

Since $CL = 0.95$, then

$$\alpha = 1 - CL = 1 - 0.95 = 0.05 \quad \frac{\alpha}{2} = 0.025.$$

Then $z_{\frac{\alpha}{2}} = z_{.025} = 1.96$

Use the TI-83, 83+ or 84+ calculator command `invNorm(0.975,0,1)` to find $z_{.025}$. Remember that the area to the right of $z_{.025}$ is 0.025 and the area to the left of $z_{.025}$ is 0.975. This can also be found using appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$EBP = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p' \cdot q'}{n}} = 1.96 \cdot \sqrt{\frac{(0.842) \cdot (0.158)}{500}} = 0.032$$

$$p' - EBP = 0.842 - 0.032 = 0.81$$

$$p' + EBP = 0.842 + 0.032 = 0.874$$

The confidence interval for the true binomial population proportion is $(p' - EBP, p' + EBP) = (0.810, 0.874)$.

Interpretation

We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have cell phones.

Explanation of 95% Confidence Level

95% of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have cell phones.

Solution:

Using a function of the TI-83, 83+ or 84 calculators:

Press **STAT** and arrow over to **TESTS**.

Arrow down to **A:1-PropZint**. Press **ENTER**.

Arrow down to x and enter 421.

Arrow down to n and enter 500.

Arrow down to **C-Level** and enter .95.

Arrow down to **Calculate** and press **ENTER**.

The confidence interval is (0.81003, 0.87397).

Example:**Exercise:****Problem:**

For a class project, a political science student at a large university wants to estimate the percent of students that are registered voters. He surveys 500 students and finds that 300 are registered voters. Compute a 90% confidence interval for the true percent of students that are registered voters and interpret the confidence interval.

Solution:

- You can use technology to directly calculate the confidence interval.
- The first solution is step-by-step (Solution A).
- The second solution uses a function of the TI-83, 83+ or 84 calculators (Solution B).

Solution A

$x = 300$ and $n = 500$.

$$p' = \frac{x}{n} = \frac{300}{500} = 0.600$$

$$q' = 1 - p' = 1 - 0.600 = 0.400$$

Since $CL = 0.90$, then

$$\alpha = 1 - CL = 1 - 0.90 = 0.10 \quad \frac{\alpha}{2} = 0.05.$$

$$z_{\frac{\alpha}{2}} = z_{.05} = 1.645$$

Use the TI-83, 83+ or 84+ calculator command `invNorm(0.95,0,1)` to find $z_{.05}$. Remember that the area to the right of $z_{.05}$ is 0.05 and the area to the left of $z_{.05}$ is 0.95. This can also be found using

appropriate commands on other calculators, using a computer, or using a Standard Normal probability table.

$$EBP = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p^* \cdot q^*}{n}} = 1.645 \cdot \sqrt{\frac{(0.60) \cdot (0.40)}{500}} = 0.036$$

$$p^* - EBP = 0.60 - 0.036 = 0.564$$

$$p^* + EBP = 0.60 + 0.036 = 0.636$$

The confidence interval for the true binomial population proportion is $(p^* - EBP, p^* + EBP) = (0.564, 0.636)$.

Interpretation:

- We estimate with 90% confidence that the true percent of all students that are registered voters is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL students are registered voters.

Explanation of 90% Confidence Level

90% of all confidence intervals constructed in this way contain the true value for the population percent of students that are registered voters.

Solution B

Using a function of the TI-83, 83+ or 84 calculators:

Press **STAT** and arrow over to **TESTS**.

Arrow down to **A:1-PropZint**. Press **ENTER**.

Arrow down to x and enter 300.

Arrow down to n and enter 500.

Arrow down to **C-Level** and enter .90.

Arrow down to **Calculate** and press **ENTER**.

The confidence interval is (0.564, 0.636).

Calculating the Sample Size n

If researchers desire a specific margin of error, then they can use the error bound formula to calculate the required sample size.

The error bound formula for a population proportion is

- $EBP = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p'q'}{n}}$
- Solving for n gives you an equation for the sample size.
- $n = \frac{z_{\frac{\alpha}{2}}^2 \cdot p'q'}{EBP^2}$

Example:

Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ that use text messaging on their cell phone. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within 3 percentage points of the true population proportion of customers aged 50+ that use text messaging on their cell phone.

Solution

From the problem, we know that **EBP=0.03** (3%=0.03) and

$z_{\frac{\alpha}{2}} = z_{.05} = 1.645$ because the confidence level is 90%

However, in order to find n , we need to know the estimated (sample) proportion p' . Remember that $q'=1-p'$. But, we do not know p' yet. Since we multiply p' and q' together, we make them both equal to 0.5 because $p'q' = (.5)(.5) = .25$ results in the largest possible product. (Try other products: $(.6)(.4) = .24$; $(.3)(.7) = .21$; $(.2)(.8) = .16$ and so on). The largest possible product gives us the largest n . This gives us a large enough sample so that we can be 90% confident that we are within 3 percentage points of the true population proportion. To calculate the sample size n , use the formula and make the substitutions.

$$n = \frac{z_{\frac{\alpha}{2}}^2 p'q'}{EBP^2} \text{ gives } n = \frac{1.645^2 (.5)(.5)}{.03^2} = 751.7$$

Round the answer to the next higher value. The sample size should be 752 cell phone customers aged 50+ in order to be 90% confident that the

estimated (sample) proportion is within 3 percentage points of the true population proportion of all customers aged 50+ that use text messaging on their cell phone.

**With contributions from Roberta Bloom.

Glossary

Binomial Distribution

A discrete random variable (RV) which arises from Bernoulli trials. There are a fixed number, n , of independent trials. “Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in n trials. The notation is: $X \sim B(n, p)$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is

$$P(X = x) = \binom{n}{x} p^x q^{n-x}.$$

Confidence Interval (CI)

An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Confidence Level (CL)

The percent expression for the probability that the confidence interval contains the true population parameter. For example, if the $CL = 90\%$, then in 90 out of 100 samples the interval estimate will enclose the true population parameter.

Error Bound for a Population Proportion(EBP)

The margin of error. Depends on the confidence level, sample size, and the estimated (from the sample) proportion of successes.

Normal Distribution

A continuous random variable (RV) with pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \text{ where } \mu \text{ is the mean of the distribution and}$$

σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Summary of Formulas

Formula General form of a confidence interval

(lower value, upper value) = (point estimate – error bound, point estimate + error bound)

Formula To find the error bound when you know the confidence interval

$$\begin{aligned} \text{error bound} &= \text{upper value} - \text{point estimate} && \text{OR} \\ \text{error bound} &= \frac{\text{upper value} - \text{lower value}}{2} \end{aligned}$$

Formula Single Population Mean, Known Standard Deviation, Normal Distribution

Use the [Normal Distribution for Means](#) $EBM = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

The confidence interval has the format ($\bar{x} - EBM$, $\bar{x} + EBM$).

Formula Single Population Mean, Unknown Standard Deviation, Student's-t Distribution

Use the Student's-t Distribution with degrees of freedom $df = n - 1$. $EBM = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$

Formula Single Population Proportion, Normal Distribution

Use the Normal Distribution for a single population proportion $p = \frac{I}{I+1}$

$$EBP = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p \cdot q}{n}} \quad I+1 = 1$$

The confidence interval has the format ($p - EBP$, $p + EBP$).

Formula Point Estimates

\bar{x} is a point estimate for μ

\hat{p} is a point estimate for p

s is a point estimate for σ

Practice 1: Confidence Intervals for Averages, Known Population Standard Deviation

Student Learning Outcomes

- The student will calculate confidence intervals for means when the population standard deviation is known.

Given

The mean age for all Foothill College students for a recent Fall term was 33.2. The population standard deviation has been pretty consistent at 15. Suppose that twenty-five Winter students were randomly selected. The mean age for the sample was 30.4. We are interested in the true mean age for Winter Foothill College students.

(http://research.fhda.edu/factbook/FH_Demo_Trends/FoothillDemographicTrends.htm)

Let X = the age of a Winter Foothill College student

Calculating the Confidence Interval

Exercise:

Problem: x

Solution:

30.4

Exercise:

Problem: $n=$

Solution:

25

Exercise:

Problem: 15=(insert symbol here)

Solution:

σ

Exercise:

Problem: Define the Random Variable, X , in words.

$X =$

Solution:

the mean age of 25 randomly selected Winter Foothill students

Exercise:

Problem: What is x estimating?

Solution:

μ

Exercise:

Problem: Is σ_x known?

Solution:

yes

Exercise:

Problem:

As a result of your answer to (4), state the exact distribution to use when calculating the Confidence Interval.

Solution:

Normal

Explaining the Confidence Interval

Construct a 95% Confidence Interval for the true mean age of Winter Foothill College students.

Exercise:

Problem: How much area is in both tails (combined)? α

Solution:

0.05

Exercise:

Problem: How much area is in each tail? $\frac{\alpha}{2}$

Solution:

0.025

Exercise:

Problem: Identify the following specifications:

- **a** lower limit =
- **b** upper limit =
- **c** error bound =

Solution:

- a24.52
- b36.28
- c5.88

Exercise:

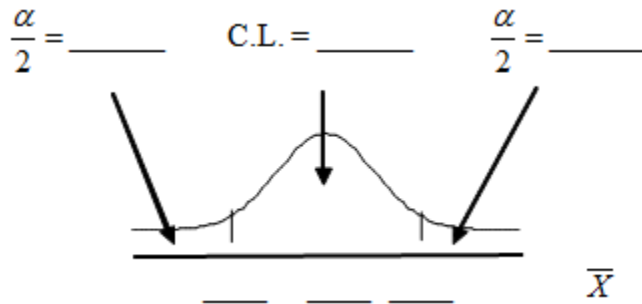
Problem: The 95% Confidence Interval is: _____

Solution:

Exercise:

Problem:

Fill in the blanks on the graph with the areas, upper and lower limits of the Confidence Interval, and the sample mean.



Exercise:

Problem: In one complete sentence, explain what the interval means.

Discussion Questions

Exercise:

Problem:

Using the same mean, standard deviation and level of confidence, suppose that n were 69 instead of 25. Would the error bound become larger or smaller? How do you know?

Exercise:**Problem:**

Using the same mean, standard deviation and sample size, how would the error bound change if the confidence level were reduced to 90%? Why?

Practice 2: Confidence Intervals for Averages, Unknown Population Standard Deviation

Student Learning Outcomes

- The student will calculate confidence intervals for means when the population standard deviation is unknown.

Given

The following real data are the result of a random survey of 39 national flags (with replacement between picks) from various countries. We are interested in finding a confidence interval for the true mean number of colors on a national flag. Let X be the number of colors on a national flag.

X	Freq.
1	1
2	7
3	18
4	7
5	6

Calculating the Confidence Interval

Exercise:

Problem: Calculate the following:

- **a**
 - **b**
 - **c**
-

Solution:

- **a**3.26
- **b**1.02
- **c**39

Exercise:

Problem:

Define the Random Variable, \bar{c} , in words.

Solution:

the mean number of colors of 39 flags

Exercise:

Problem: What is \bar{c} estimating?

Solution:

Exercise:

Problem: Is \bar{c} known?

Solution:

No

Exercise:

Problem:

As a result of your answer to (4), state the exact distribution to use when calculating the Confidence Interval.

Solution:

Confidence Interval for the True Mean Number

Construct a 95% Confidence Interval for the true mean number of colors on national flags.

Exercise:

Problem: How much area is in both tails (combined)?

Solution:

0.05

Exercise:

Problem: How much area is in each tail? —

Solution:

0.025

Exercise:

Problem: Calculate the following:

- $\alpha_{\text{lower limit}} =$

- upper limit =
- error bound =

Solution:

- a 2.93
- b 3.59
- c 0.33

Exercise:

Problem: The 95% Confidence Interval is:

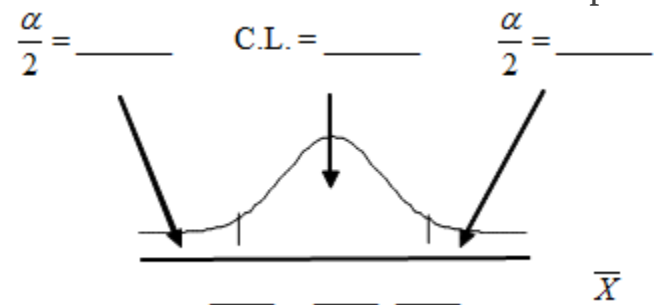
Solution:

2.93; 3.59

Exercise:

Problem:

Fill in the blanks on the graph with the areas, upper and lower limits of the Confidence Interval and the sample mean.



Exercise:

Problem: In one complete sentence, explain what the interval means.

Discussion Questions

Exercise:**Problem:**

Using the same σ , n , and level of confidence, suppose that \bar{x} were 69 instead of 39. Would the error bound become larger or smaller? How do you know?

Exercise:**Problem:**

Using the same σ , n , and \bar{x} , how would the error bound change if the confidence level were reduced to 90%? Why?

Practice 3: Confidence Intervals for Proportions

Student Learning Outcomes

- The student will calculate confidence intervals for proportions.

Given

The Ice Chalet offers dozens of different beginning ice-skating classes. All of the class names are put into a bucket. The 5 P.M., Monday night, ages 8 - 12, beginning ice-skating class was picked. In that class were 64 girls and 16 boys. Suppose that we are interested in the true proportion of girls, ages 8 - 12, in all beginning ice-skating classes at the Ice Chalet. Assume that the children in the selected class is a random sample of the population.

Estimated Distribution

Exercise:

Problem: What is being counted?

Exercise:

Problem: In words, define the Random Variable X . $X =$

Solution:

The number of girls, age 8-12, in the beginning ice skating class

Exercise:

Problem: Calculate the following:

- **a** $x =$
- **b** $n =$
- **c** $p =$

Solution:

- a64
- b80
- c0.8

Exercise:

Problem: State the estimated distribution of X . $X \sim$

Solution:

$B(80, 0.80)$

Exercise:

Problem: Define a new Random Variable P_l . What is p_l estimating?

Solution:

p

Exercise:

Problem: In words, define the Random Variable P_l . $P_l =$

Solution:

The proportion of girls, age 8-12, in the beginning ice skating class.

Exercise:

Problem: State the estimated distribution of P_l . $P_l \sim$

Explaining the Confidence Interval

Construct a 92% Confidence Interval for the true proportion of girls in the age 8 - 12 beginning ice-skating classes at the Ice Chalet.

Exercise:

Problem: How much area is in both tails (combined)? $\alpha =$

Solution:

$$1 - 0.92 = 0.08$$

Exercise:

Problem: How much area is in each tail? $\frac{\alpha}{2} =$

Solution:

$$0.04$$

Exercise:

Problem: Calculate the following:

- a lower limit =
- b upper limit =
- c error bound =

Solution:

- a 0.72
- b 0.88
- c 0.08

Exercise:

Problem: The 92% Confidence Interval is:

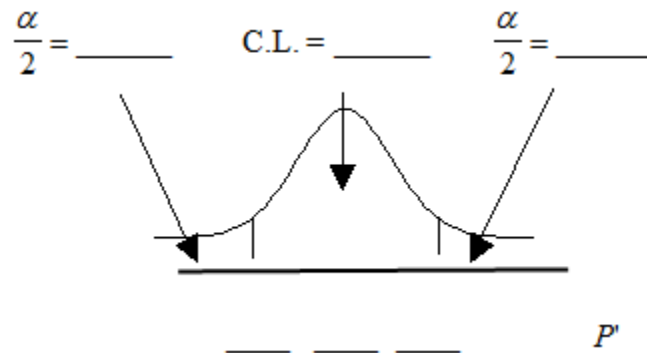
Solution:

(0.72; 0.88)

Exercise:

Problem:

Fill in the blanks on the graph with the areas, upper and lower limits of the Confidence Interval, and the sample proportion.



Exercise:

Problem: In one complete sentence, explain what the interval means.

Discussion Questions

Exercise:

Problem:

Using the same p' and level of confidence, suppose that n were increased to 100. Would the error bound become larger or smaller? How do you know?

Exercise:

Problem:

Using the same p' and $n = 80$, how would the error bound change if the confidence level were increased to 98%? Why?

Exercise:

Problem:

If you decreased the allowable error bound, why would the minimum sample size increase (keeping the same level of confidence)?

Homework

Note: If you are using a student's-t distribution for a homework problem below, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, though.)

Exercise:

Problem:

Among various ethnic groups, the standard deviation of heights is known to be approximately 3 inches. We wish to construct a 95% confidence interval for the mean height of male Swedes. 48 male Swedes are surveyed. The sample mean is 71 inches. The sample standard deviation is 2.8 inches.

- **a**
 - **i** $\bar{x} =$ _____
 - **ii** $\sigma =$ _____
 - **iii** $s_x =$ _____
 - **iv** $n =$ _____
 - **v** $n - 1 =$ _____
- **b** Define the Random Variables X and \bar{X} , in words.
- **c** Which distribution should you use for this problem? Explain your choice.
- **d** Construct a 95% confidence interval for the population mean height of male Swedes.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.

- **e**What will happen to the level of confidence obtained if 1000 male Swedes are surveyed instead of 48? Why?
-

Solution:

- **a**
 - **i**71
 - **ii**3
 - **iii**2.8
 - **iv**48
 - **v**47
- **c** $N\left(71, \frac{3}{\sqrt{48}}\right)$
- **d**
 - **i**CI: (70.15,71.85)
 - **iii**EB = 0.85

Exercise:

Problem:

In six packages of “The Flintstones® Real Fruit Snacks” there were 5 Bam-Bam snack pieces. The total number of snack pieces in the six bags was 68. We wish to calculate a 96% confidence interval for the population proportion of Bam-Bam snack pieces.

- **a**Define the Random Variables X and P' , in words.
- **b**Which distribution should you use for this problem? Explain your choice
- **c**Calculate p' .
- **d**Construct a 96% confidence interval for the population proportion of Bam-Bam snack pieces per bag.
 - **i** State the confidence interval.
 - **ii**Sketch the graph.

- **iii** Calculate the error bound.
- **e** Do you think that six packages of fruit snacks yield enough data to give accurate results? Why or why not?

Exercise:

Problem:

A random survey of enrollment at 35 community colleges across the United States yielded the following figures (source: Microsoft Bookshelf): 6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622. Assume the underlying population is normal.

- **a**
 - **i** $\bar{x} =$
 - **ii** $s_x =$ _____
 - **iii** $n =$ _____
 - **iv** $n - 1 =$ _____
- **b** Define the Random Variables X and \bar{X} , in words.
- **c** Which distribution should you use for this problem? Explain your choice.
- **d** Construct a 95% confidence interval for the population mean enrollment at community colleges in the United States.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.
- **e** What will happen to the error bound and confidence interval if 500 community colleges were surveyed? Why?

Solution:

- **a**
 - **i**8629
 - **ii**6944
 - **iii**35
 - **iv**34
- **c** t_{34}
- **d**
 - **i**CI: (6244, 11,014)
 - **iii**EB = 2385
- **e**It will become smaller

Exercise:**Problem:**

From a stack of IEEE Spectrum magazines, announcements for 84 upcoming engineering conferences were randomly picked. The mean length of the conferences was 3.94 days, with a standard deviation of 1.28 days. Assume the underlying population is normal.

- **a** Define the Random Variables X and \bar{X} , in words.
- **b** Which distribution should you use for this problem? Explain your choice.
- **c** Construct a 95% confidence interval for the population mean length of engineering conferences.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.

Exercise:

Problem:

Suppose that a committee is studying whether or not there is waste of time in our judicial system. It is interested in the mean amount of time individuals waste at the courthouse waiting to be called for service. The committee randomly surveyed 81 people. The sample mean was 8 hours with a sample standard deviation of 4 hours.

- **a**
 - **i** $\bar{x} =$ _____
 - **ii** $s_x =$ _____
 - **iii** $n =$ _____
 - **iv** $n - 1 =$ _____
 - **b** Define the Random Variables X and \bar{X} , in words.
 - **c** Which distribution should you use for this problem? Explain your choice.
 - **d** Construct a 95% confidence interval for the population mean time wasted.
 - **a** State the confidence interval.
 - **b** Sketch the graph.
 - **c** Calculate the error bound.
 - **e** Explain in a complete sentence what the confidence interval means.
-

Solution:

- **a**
 - **i** 8
 - **ii** 4
 - **iii** 81
 - **iv** 80

- **c** t_{80}
- **d**
 - **i**CI: (7.12, 8.88)
 - **iii**EB = 0.88

Exercise:

Problem:

Suppose that an accounting firm does a study to determine the time needed to complete one person's tax forms. It randomly surveys 100 people. The sample mean is 23.6 hours. There is a known standard deviation of 7.0 hours. The population distribution is assumed to be normal.

- **a**
 - **i** $x =$ _____
 - **ii** $\sigma =$ _____
 - **iii** $s_x =$ _____
 - **iv** $n =$ _____
 - **v** $n - 1 =$ _____
- **b** Define the Random Variables X and \bar{X} , in words.
- **c** Which distribution should you use for this problem? Explain your choice.
- **d** Construct a 90% confidence interval for the population mean time to complete the tax forms.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.
- **e** If the firm wished to increase its level of confidence and keep the error bound the same by taking another survey, what changes should it make?

- **f**If the firm did another survey, kept the error bound the same, and only surveyed 49 people, what would happen to the level of confidence? Why?
- **g**Suppose that the firm decided that it needed to be at least 96% confident of the population mean length of time to within 1 hour. How would the number of people the firm surveys change? Why?

Exercise:

Problem:

A sample of 16 small bags of the same brand of candies was selected. Assume that the population distribution of bag weights is normal. The weight of each bag was then recorded. The mean weight was 2 ounces with a standard deviation of 0.12 ounces. The population standard deviation is known to be 0.1 ounce.

- **a**
 - **i** $x =$ _____
 - **ii** $\sigma =$ _____
 - **iii** $s_x =$ _____
 - **iv** $n =$ _____
 - **v** $n - 1 =$ _____
- **b**Define the Random Variable X , in words.
- **c**Define the Random Variable X , in words.
- **d**Which distribution should you use for this problem? Explain your choice.
- **e**Construct a 90% confidence interval for the population mean weight of the candies.
 - **i**State the confidence interval.
 - **ii**Sketch the graph.
 - **iii**Calculate the error bound.
- **f**Construct a 98% confidence interval for the population mean weight of the candies.

- **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.
 - **g** In complete sentences, explain why the confidence interval in (f) is larger than the confidence interval in (e).
 - **h** In complete sentences, give an interpretation of what the interval in (f) means.
-

Solution:

- **a**
 - **i** 2
 - **ii** 0.1
 - **iii** 0.12
 - **iv** 16
 - **v** 15
- **b** the weight of 1 small bag of candies
- **c** the mean weight of 16 small bags of candies
- **d** $N\left(2, \frac{0.1}{\sqrt{16}}\right)$
- **e**
 - **i** CI: (1.96, 2.04)
 - **iii** EB = 0.04
- **f**
 - **i** CI: (1.94, 2.06)
 - **iii** EB = 0.06

Exercise:

Problem:

A pharmaceutical company makes tranquilizers. It is assumed that the distribution for the length of time they last is approximately normal. Researchers in a hospital used the drug on a random sample of 9 patients. The effective period of the tranquilizer for each patient (in hours) was as follows: 2.7; 2.8; 3.0; 2.3; 2.3; 2.2; 2.8; 2.1; and 2.4 .

- **a**
 - **i** $\bar{x} =$ _____
 - **ii** $s_x =$ _____
 - **iii** $n =$ _____
 - **iv** $n - 1 =$ _____
- **b** Define the Random Variable X , in words.
- **c** Define the Random Variable X , in words.
- **d** Which distribution should you use for this problem? Explain your choice.
- **e** Construct a 95% confidence interval for the population mean length of time.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.
- **f** What does it mean to be “95% confident” in this problem?

Exercise:**Problem:**

Suppose that 14 children were surveyed to determine how long they had to use training wheels. It was revealed that they used them an average of 6 months with a sample standard deviation of 3 months. Assume that the underlying population distribution is normal.

- **a**

- **i** $x =$ _____
 - **ii** $s_x =$ _____
 - **iii** $n =$ _____
 - **iv** $n - 1 =$ _____
 - **b** Define the Random Variable X , in words.
 - **c** Define the Random Variable X , in words.
 - **d** Which distribution should you use for this problem? Explain your choice.
 - **e** Construct a 99% confidence interval for the population mean length of time using training wheels.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.
 - **f** Why would the error bound change if the confidence level was lowered to 90%?
-

Solution:

- **a**
 - **i** 6
 - **ii** 3
 - **iii** 14
 - **iv** 13
- **b** the time for a child to remove his training wheels
- **c** the mean time for 14 children to remove their training wheels.
- **d** t_{13}
- **e**
 - **i** CI: (3.58, 8.42)
 - **iii** EB = 2.42

Exercise:

Problem:

Insurance companies are interested in knowing the population percent of drivers who always buckle up before riding in a car.

- **a** When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 95% confident that the population proportion is estimated to within 0.03?
- **b** If it was later determined that it was important to be more than 95% confident and a new survey was commissioned, how would that affect the minimum number you would need to survey? Why?

Exercise:

Problem:

Suppose that the insurance companies did do a survey. They randomly surveyed 400 drivers and found that 320 claimed to always buckle up. We are interested in the population proportion of drivers who claim to always buckle up.

- **a**
 - **i** $x =$ _____
 - **ii** $n =$ _____
 - **iii** $p' =$ _____
- **b** Define the Random Variables X and P' , in words.
- **c** Which distribution should you use for this problem? Explain your choice.
- **d** Construct a 95% confidence interval for the population proportion that claim to always buckle up.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.

- **e** If this survey were done by telephone, list 3 difficulties the companies might have in obtaining random results.

Solution:

- **a**
 - **i** 320
 - **ii** 400
 - **iii** 0.80
- **c** $N\left(0.80, \sqrt{\frac{(0.80)(0.20)}{400}}\right)$
- **d**
 - **i** CI: (0.76, 0.84)
 - **iii** EB = 0.04

Exercise:

Problem:

Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its mean number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected and the number of unoccupied seats is noted for each of the sampled flights. The sample mean is 11.6 seats and the sample standard deviation is 4.1 seats.

- **a**
 - **i** $x = \underline{\hspace{2cm}}$
 - **ii** $s_x = \underline{\hspace{2cm}}$
 - **iii** $n = \underline{\hspace{2cm}}$
 - **iv** $n - 1 = \underline{\hspace{2cm}}$
- **b** Define the Random Variables X and \bar{X} , in words.

- **c** Which distribution should you use for this problem? Explain your choice.
- **d** Construct a 92% confidence interval for the population mean number of unoccupied seats per flight.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.

Exercise:

Problem:

According to a recent survey of 1200 people, 61% feel that the president is doing an acceptable job. We are interested in the population proportion of people who feel the president is doing an acceptable job.

- **a** Define the Random Variables X and P' , in words.
- **b** Which distribution should you use for this problem? Explain your choice.
- **c** Construct a 90% confidence interval for the population proportion of people who feel the president is doing an acceptable job.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.

Solution:

- **b** $N\left(0.61, \sqrt{\frac{(0.61)(0.39)}{1200}}\right)$
- **c**
 - **i** CI: (0.59, 0.63)
 - **iii** EB = 0.02

Exercise:

Problem:

A survey of the mean amount of cents off that coupons give was done by randomly surveying one coupon per page from the coupon sections of a recent San Jose Mercury News. The following data were collected: 20¢; 75¢; 50¢; 65¢; 30¢; 55¢; 40¢; 40¢; 30¢; 55¢; \$1.50; 40¢; 65¢; 40¢. Assume the underlying distribution is approximately normal.

- **a**
 - **i** $\bar{x} =$ _____
 - **ii** $s_x =$ _____
 - **iii** $n =$ _____
 - **iv** $n - 1 =$ _____
- **b** Define the Random Variables X and \bar{X} , in words.
- **c** Which distribution should you use for this problem? Explain your choice.
- **d** Construct a 95% confidence interval for the population mean worth of coupons.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.
- **e** If many random samples were taken of size 14, what percent of the confident intervals constructed should contain the population mean worth of coupons? Explain why.

Exercise:

Problem:

An article regarding interracial dating and marriage recently appeared in the Washington Post. Of the 1709 randomly selected adults, 315 identified themselves as Latinos, 323 identified themselves as blacks, 254 identified themselves as Asians, and 779 identified themselves as whites. In this survey, 86% of blacks said that their families would welcome a white person into their families. Among Asians, 77% would welcome a white person into their families, 71% would welcome a Latino, and 66% would welcome a black person.

- **a** We are interested in finding the 95% confidence interval for the percent of all black families that would welcome a white person into their families. Define the Random Variables X and P' , in words.
- **b** Which distribution should you use for this problem? Explain your choice.
- **c** Construct a 95% confidence interval
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.

Solution:

- **b** $N\left(0.86, \sqrt{\frac{(0.86)(0.14)}{323}}\right)$
- **c**
 - **i** CI: (0.823, 0.898)
 - **iii** EB = 0.038

Exercise:

Problem: Refer to the problem [above](#).

- **a** Construct three 95% confidence intervals.
 - **i** Percent of all Asians that would welcome a white person into their families.
 - **ii** Percent of all Asians that would welcome a Latino into their families.
 - **iii** Percent of all Asians that would welcome a black person into their families.
- **b** Even though the three point estimates are different, do any of the confidence intervals overlap? Which?
- **c** For any intervals that do overlap, in words, what does this imply about the significance of the differences in the true proportions?
- **d** For any intervals that do not overlap, in words, what does this imply about the significance of the differences in the true proportions?

Exercise:

Problem:

A camp director is interested in the mean number of letters each child sends during his/her camp session. The population standard deviation is known to be 2.5. A survey of 20 campers is taken. The mean from the sample is 7.9 with a sample standard deviation of 2.8.

- **a**
 - **i** $x =$ _____
 - **ii** $\sigma =$ _____
 - **iii** $s_x =$ _____
 - **iv** $n =$ _____
 - **v** $n - 1 =$ _____
- **b** Define the Random Variables X and \bar{X} , in words.
- **c** Which distribution should you use for this problem? Explain your choice.

- **d** Construct a 90% confidence interval for the population mean number of letters campers send home.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.
 - **e** What will happen to the error bound and confidence interval if 500 campers are surveyed? Why?
-

Solution:

- **a**
 - **i** 7.9
 - **ii** 2.5
 - **iii** 2.8
 - **iv** 20
 - **v** 19
- **c** $N(7.9, \frac{2.5}{\sqrt{20}})$
- **d**
 - **i** CI: (6.98, 8.82)
 - **iii** EB: 0.92

Exercise:

Problem:

Stanford University conducted a study of whether running is healthy for men and women over age 50. During the first eight years of the study, 1.5% of the 451 members of the 50-Plus Fitness Association died. We are interested in the proportion of people over 50 who ran and died in the same eight-year period.

- **a** Define the Random Variables X and P' , in words.

- **b** Which distribution should you use for this problem? Explain your choice.
- **c** Construct a 97% confidence interval for the population proportion of people over 50 who ran and died in the same eight-year period.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.
- **d** Explain what a “97% confidence interval” means for this study.

Exercise:

Problem:

In a recent sample of 84 used cars sales costs, the sample mean was \$6425 with a standard deviation of \$3156. Assume the underlying distribution is approximately normal.

- **a** Which distribution should you use for this problem? Explain your choice.
- **b** Define the Random Variable X , in words.
- **c** Construct a 95% confidence interval for the population mean cost of a used car.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.
- **d** Explain what a “95% confidence interval” means for this study.

Solution:

- **a** t_{83}
- **b** mean cost of 84 used cars
- **c**

- **i**CI: (5740.10, 7109.90)
- **iii** EB = 684.90

Exercise:

Problem:

A telephone poll of 1000 adult Americans was reported in an issue of Time Magazine. One of the questions asked was “What is the main problem facing the country?” 20% answered “crime”. We are interested in the population proportion of adult Americans who feel that crime is the main problem.

- **a** Define the Random Variables X and P' , in words.
- **b** Which distribution should you use for this problem? Explain your choice.
- **c** Construct a 95% confidence interval for the population proportion of adult Americans who feel that crime is the main problem.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.
- **d** Suppose we want to lower the sampling error. What is one way to accomplish that?
- **e** The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is $\pm 3\%$. In 1-3 complete sentences, explain what the $\pm 3\%$ represents.

Exercise:

Problem:

Refer to the above problem. Another question in the poll was “[How much are] you worried about the quality of education in our schools?” 63% responded “a lot”. We are interested in the population proportion of adult Americans who are worried a lot about the quality of education in our schools.

1. Define the Random Variables X and P' , in words.
2. Which distribution should you use for this problem? Explain your choice.
3. Construct a 95% confidence interval for the population proportion of adult Americans worried a lot about the quality of education in our schools.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.
4. The sampling error given by Yankelovich Partners, Inc. (which conducted the poll) is $\pm 3\%$. In 1-3 complete sentences, explain what the $\pm 3\%$ represents.

Solution:

- **b** $N\left(0.63, \sqrt{\frac{(0.63)(0.37)}{1000}}\right)$
- **c**
 - **i** CI: (0.60, 0.66)
 - **iii** EB = 0.03

Exercise:

Problem:

Six different national brands of chocolate chip cookies were randomly selected at the supermarket. The grams of fat per serving are as follows: 8; 8; 10; 7; 9; 9. Assume the underlying distribution is approximately normal.

- **a** Calculate a 90% confidence interval for the population mean grams of fat per serving of chocolate chip cookies sold in supermarkets.
 - **i** State the confidence interval.
 - **ii** Sketch the graph.
 - **iii** Calculate the error bound.
- **b** If you wanted a smaller error bound while keeping the same level of confidence, what should have been changed in the study before it was done?
- **c** Go to the store and record the grams of fat per serving of six brands of chocolate chip cookies.
- **d** Calculate the mean.
- **e** Is the mean within the interval you calculated in part (a)? Did you expect it to be? Why or why not?

Exercise:**Problem:**

A confidence interval for a proportion is given to be $(-0.22, 0.34)$. Why doesn't the lower limit of the confidence interval make practical sense? How should it be changed? Why?

Try these multiple choice questions.

The next three problems refer to the following: According to a Field Poll, 79% of California adults (actual results are 400 out of 506 surveyed) feel that “education and our schools” is one of the top issues facing

California. We wish to construct a 90% confidence interval for the true proportion of California adults who feel that education and the schools is one of the top issues facing California. (Source: <http://field.com/fieldpollonline/subscribers/>)

Exercise:

Problem:A point estimate for the true population proportion is:

- A 0.90
- B 1.27
- C 0.79
- D 400

Solution:

C

Exercise:

Problem:A 90% confidence interval for the population proportion is:

- A (0.761, 0.820)
- B (0.125, 0.188)
- C (0.755, 0.826)
- D (0.130, 0.183)

Solution:

A

Exercise:

Problem:The error bound is approximately

- A 1.581
- B 0.791

- C0.059
- D0.030

Solution:

D

The next two problems refer to the following:

A quality control specialist for a restaurant chain takes a random sample of size 12 to check the amount of soda served in the 16 oz. serving size. The sample mean is 13.30 with a sample standard deviation of 1.55. Assume the underlying population is normally distributed.

Exercise:

Problem:

Find the 95% Confidence Interval for the true population mean for the amount of soda served.

- A(12.42, 14.18)
- B(12.32, 14.29)
- C(12.50, 14.10)
- DImpossible to determine

Solution:

B

Exercise:

Problem:What is the error bound?

- A0.87
- B1.98
- C0.99
- D1.74

Solution:

C

Exercise:**Problem:**

What is meant by the term “90% confident” when constructing a confidence interval for a mean?

- **A**If we took repeated samples, approximately 90% of the samples would produce the same confidence interval.
- **B**If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the sample mean.
- **C**If we took repeated samples, approximately 90% of the confidence intervals calculated from those samples would contain the true value of the population mean.
- **D**If we took repeated samples, the sample mean would equal the population mean in approximately 90% of the samples.

Solution:

C

The next two problems refer to the following:

Five hundred and eleven (511) homes in a certain southern California community are randomly surveyed to determine if they meet minimal earthquake preparedness recommendations. One hundred seventy-three (173) of the homes surveyed met the minimum recommendations for earthquake preparedness and 338 did not.

Exercise:

Problem:

Find the Confidence Interval at the 90% Confidence Level for the true population proportion of southern California community homes meeting at least the minimum recommendations for earthquake preparedness.

- A(0.2975, 0.3796)
 - B(0.6270, 0.6959)
 - C(0.3041, 0.3730)
 - D(0.6204, 0.7025)
-

Solution:

C

Exercise:**Problem:**

The point estimate for the population proportion of homes that do not meet the minimum recommendations for earthquake preparedness is:

- A0.6614
 - B0.3386
 - C173
 - D338
-

Solution:

A

Review

The next three problems refer to the following situation: Suppose that a sample of 15 randomly chosen people were put on a special weight loss diet. The amount of weight lost, in pounds, follows an unknown distribution with mean equal to 12 pounds and standard deviation equal to 3 pounds. Assume that the distribution for the weight loss is normal.

Exercise:

Problem:

To find the probability that the mean amount of weight lost by 15 people is no more than 14 pounds, the random variable should be:

- **A** The number of people who lost weight on the special weight loss diet
- **B** The number of people who were on the diet
- **C** The mean amount of weight lost by 15 people on the special weight loss diet
- **D** The total amount of weight lost by 15 people on the special weight loss diet

Solution:

C

Exercise:

Problem: Find the probability asked for in the previous problem.

Solution:

0.9951

Exercise:

Problem:

Find the 90th percentile for the mean amount of weight lost by 15 people.

Solution:

12.99

The next three questions refer to the following situation: The time of occurrence of the first accident during rush-hour traffic at a major intersection is uniformly distributed between the three hour interval 4 p.m. to 7 p.m. Let T = the amount of time (hours) it takes for the first accident to occur.

- So, if an accident occurs at 4 p.m., the amount of time, in hours, it took for the accident to occur is _____.
- _____
- _____

Exercise:**Problem:**

What is the probability that the time of occurrence is within the first half-hour or the last hour of the period from 4 to 7 p.m.?

- A Cannot be determined from the information given
- B —
- C —
- D —

Solution:

C

Exercise:

Problem: The 20th percentile occurs after how many hours?

- A 0.20
 - B 0.60
 - C 0.50
 - D 1
-

Solution:

B

Exercise:

Problem:

Assume Ramon has kept track of the times for the first accidents to occur for 40 different days. Let T = the total cumulative time. Then follows which distribution?

- A
 - B —
 - C
 - D
-

Solution:

C

Exercise:

Problem:

Using the information in question #6, find the probability that the total time for all first accidents to occur is more than 43 hours.

Solution:

0.9990

The next two questions refer to the following situation: The length of time a parent must wait for his children to clean their rooms is uniformly distributed in the time interval from 1 to 15 days.

Exercise:

Problem:

How long must a parent expect to wait for his children to clean their rooms?

- A 8 days
- B 3 days
- C 14 days
- D 6 days

Solution:

A

Exercise:

Problem:

What is the probability that a parent will wait more than 6 days given that the parent has already waited more than 3 days?

- A 0.5174
- B 0.0174
- C 0.7500
- D 0.2143

Solution:

C

The next five problems refer to the following study: Twenty percent of the students at a local community college live in within five miles of the campus. Thirty percent of the students at the same community college receive some kind of financial aid. Of those who live within five miles of the campus, 75% receive some kind of financial aid.

Exercise:

Problem:

Find the probability that a randomly chosen student at the local community college does not live within five miles of the campus.

- A 80%
 - B 20%
 - C 30%
 - D Cannot be determined
-

Solution:

A

Exercise:

Problem:

Find the probability that a randomly chosen student at the local community college lives within five miles of the campus or receives some kind of financial aid.

- A 50%
 - B 35%
 - C 27.5%
 - D 75%
-

Solution:

B

Exercise:**Problem:**

Based upon the above information, are living in student housing within five miles of the campus and receiving some kind of financial aid mutually exclusive?

- **A** Yes
- **B** No
- **C** Cannot be determined

Solution:

B

Exercise:**Problem:**

The interest rate charged on the financial aid is _____ data.

- **A** quantitative discrete
- **B** quantitative continuous
- **C** qualitative discrete
- **D** qualitative

Solution:

B

Exercise:**Problem:**

What follows is information about the students who receive financial aid at the local community college.

- 1st quartile = \$250

- 2nd quartile = \$700
- 3rd quartile = \$1200

(These amounts are for the school year.) If a sample of 200 students is taken, how many are expected to receive \$250 or more?

- **A** 50
- **B** 250
- **C** 150
- **D** Cannot be determined

Solution:

- **C** 150

The next two problems refer to the following information: , , and are independent events.

Exercise:

Problem:

- **A** 0.5
- **B** 0.6
- **C** 0
- **D** 0.06

Solution:

D

Exercise:

Problem:

- **A** 0.56

- **B** 0.5
- **C** 0.44
- **D** 1

Solution:

C

Exercise:

Problem:

If A and B are mutually exclusive events, $P(A) = 0.5$, $P(B) = 0.4$, then $P(A \cap B) =$

- **A** 1
- **B** 0
- **C** 0.40
- **D** 0.0375

Solution:

B

Lab 1: Confidence Interval (Home Costs)

Class Time:

Names:

Student Learning Outcomes:

- The student will calculate the 90% confidence interval for the mean cost of a home in the area in which this school is located.
- The student will interpret confidence intervals.
- The student will determine the effects that changing conditions has on the confidence interval.

Collect the Data

Check the Real Estate section in your local newspaper. (Note: many papers only list them one day per week. Also, we will assume that homes come up for sale randomly.) Record the sales prices for 35 randomly selected homes recently listed in the county.

1. Complete the table:

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

Describe the Data

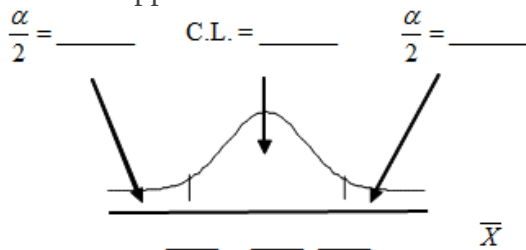
1. Compute the following:

- **a** $x =$
- **b** $s_x =$
- **c** $n =$

2. Define the Random Variable X , in words. $X =$
3. State the estimated distribution to use. Use both words and symbols.

Find the Confidence Interval

1. Calculate the confidence interval and the error bound.
 - **a** Confidence Interval:
 - **b** Error Bound:
2. How much area is in both tails (combined)? $\alpha =$
3. How much area is in each tail? $\frac{\alpha}{2} =$
4. Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample mean.



5. Some students think that a 90% confidence interval contains 90% of the data. Use the list of data on the first page and count how many of the data values lie within the confidence interval. What percent is this? Is this percent close to 90%? Explain why this percent should or should not be close to 90%.

Describe the Confidence Interval

1. In two to three complete sentences, explain what a Confidence Interval means (in general), as if you were talking to someone who has not taken statistics.
2. In one to two complete sentences, explain what this Confidence Interval means for this particular study.

Use the Data to Construct Confidence Intervals

1. Using the above information, construct a confidence interval for each confidence level given.

Confidence level	EBM / Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

2. What happens to the EBM as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

Lab 2: Confidence Interval (Place of Birth)

Class Time:

Names:

Student Learning Outcomes:

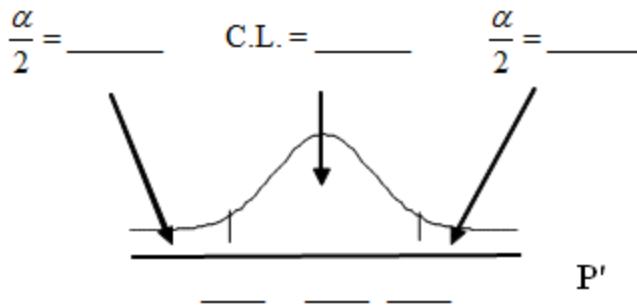
- The student will calculate the 90% confidence interval for proportion of students in this school that were born in this state.
- The student will interpret confidence intervals.
- The student will determine the effects that changing conditions have on the confidence interval.

Collect the Data

1. Survey the students in your class, asking them if they were born in this state. Let X = the number that were born in this state.
 - **a** n = _____
 - **b** x = _____
2. Define the Random Variable P' in words.
3. State the estimated distribution to use.

Find the Confidence Interval and Error Bound

1. Calculate the confidence interval and the error bound.
 - **a** Confidence Interval:
 - **b** Error Bound:
2. How much area is in both tails (combined)? α =
3. How much area is in each tail? $\frac{\alpha}{2}$ =
4. Fill in the blanks on the graph with the area in each section. Then, fill in the number line with the upper and lower limits of the confidence interval and the sample proportion.



Describe the Confidence Interval

1. In two to three complete sentences, explain what a Confidence Interval means (in general), as if you were talking to someone who has not taken statistics.
2. In one to two complete sentences, explain what this Confidence Interval means for this particular study.
3. Using the above information, construct a confidence interval for each given confidence level given.

Confidence level	EBP / Error Bound	Confidence Interval
50%		
80%		
95%		
99%		

4. What happens to the EBP as the confidence level increases? Does the width of the confidence interval increase or decrease? Explain why this happens.

Lab 3: Confidence Interval (Womens' Heights)

Class Time:

Names:

Student Learning Outcomes:

- The student will calculate a 90% confidence interval using the given data.
- The student will determine the relationship between the confidence level and the percent of constructed intervals that contain the population mean.

Given:

1.	59.4	71.6	69.3	65.0	62.9
	66.5	61.7	55.2	67.5	67.2
	63.8	62.9	63.0	63.9	68.7
	65.5	61.9	69.6	58.7	63.4
	61.8	60.6	69.8	60.0	64.9
	66.1	66.8	60.6	65.6	63.8
	61.3	59.2	64.1	59.3	64.9
	62.4	63.5	60.9	63.3	66.3
	61.5	64.3	62.9	60.6	63.8
	58.8	64.9	65.7	62.5	70.9
	62.9	63.1	62.2	58.7	64.7
	66.0	60.5	64.7	65.4	60.2
	65.0	64.1	61.1	65.3	64.6

90% Confidence Intervals

Discussion Questions

1. The actual population mean for the 100 heights given above is $\mu = 63.4$. Using the class listing of confidence intervals, count how many of them contain the population mean μ ; i.e., for how many intervals does the value of μ lie between the endpoints of the confidence interval?
2. Divide this number by the total number of confidence intervals generated by the class to determine the percent of confidence intervals that contains the mean μ . Write this percent below.
3. Is the percent of confidence intervals that contain the population mean μ close to 90%?
4. Suppose we had generated 100 confidence intervals. What do you think would happen to the percent of confidence intervals that contained the population mean?
5. When we construct a 90% confidence interval, we say that we are **90% confident that the true population mean lies within the confidence interval**. Using complete sentences, explain what we mean by this phrase.
6. Some students think that a 90% confidence interval contains 90% of the data. Use the list of data given (the heights of women) and count how many of the data values lie within the confidence interval that you generated on that page. How many of the 100 data values lie within your confidence interval? What percent is this? Is this percent close to 90%?
7. Explain why it does not make sense to count data values that lie in a confidence interval. Think about the random variable that is being used in the problem.
8. Suppose you obtained the heights of 10 women and calculated a confidence interval from this information. Without knowing the population mean μ , would you have any way of knowing **for certain** if your interval actually contained the value of μ ? Explain.

Note: This lab was designed and contributed by Diane Mathios.

Introduction to Bivariate Data

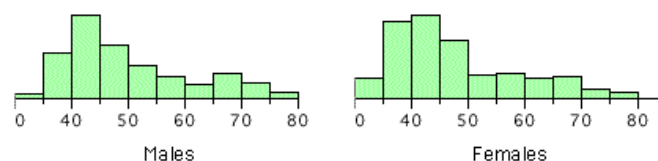
Measures of central tendency, variability, and spread summarize a single variable by providing important information about its distribution. Often, more than one variable is collected on each individual. For example, in large health studies of populations it is common to obtain variables such as age, sex, height, weight, blood pressure, and total cholesterol on each individual. Economic studies may be interested in, among other things, personal income and years of education. As a third example, most university admissions committees ask for an applicant's high school grade point average and standardized admission test scores (e.g., SAT). In this chapter we consider bivariate data, which for now consists of two [quantitative variables](#) for each individual. Our first interest is in summarizing such data in a way that is analogous to summarizing univariate (single variable) data.

By way of illustration, let's consider something with which we are all familiar: age. It helps to discuss something familiar since knowing the subject matter goes a long way in making judgments about statistical results. Let's begin by asking if people tend to marry other people of about the same age. Our experience tells us "yes," but how good is the correspondence? One way to address the question is to look at pairs of ages for a sample of married couples. Table 1 below shows the ages of 10 married couples. Going across the columns we see that, yes, husbands and wives tend to be of about the same age, with men having a tendency to be slightly older than their wives. This is no big surprise, but at least the data bear out our experiences, which is not always the case.

Husband	36	72	37	36	51	50	47	50	37	41
Wife	35	67	33	35	50	46	47	42	36	41

Sample of spousal ages of 10 White American Couples.

The pairs of ages in [\[link\]](#) are from a dataset consisting of 282 pairs of spousal ages, too many to make sense of from a table. What we need is a way to summarize the 282 pairs of ages. We know that each variable can be summarized by a [histogram](#) (see [\[link\]](#)) and by a mean and standard deviation (See [\[link\]](#)).



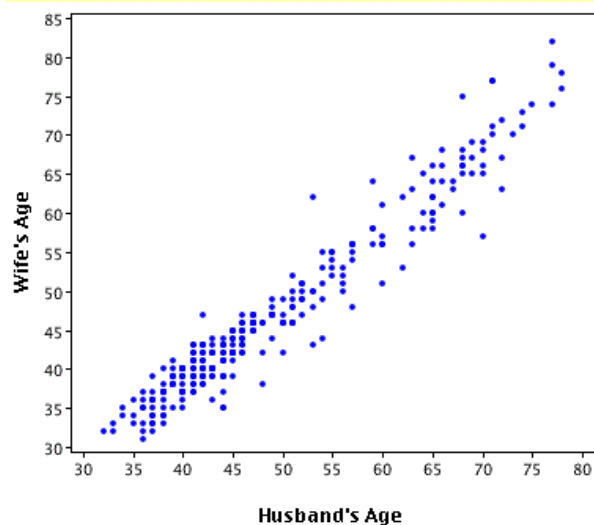
Histograms of spousal ages.

	Mean	Standard Deviation
Husband	49	11
Wife	47	11

Means and standard deviations of spousal ages.

Each distribution is fairly skewed with a long right tail. From [\[link\]](#) we see that not all husbands are older than their wives and it is important to see that this fact is lost when we separate the variables. That is, even though we provide summary statistics on each variable, the pairing within couple is lost by separating the variables. We cannot say, for example, based on the means alone what percentage of couples have younger husbands than wives. We have to count across pairs to find this out. Only by maintaining the pairing can meaningful answers be found about couples per se. Another example of information not available from the separate descriptions of husbands and wives' ages is the mean age of husbands with wives of a certain age. For instance, what is the average age of husbands with 45-year-old wives? Finally, we do not know the relationship between the husband's age and the wife's age.

We can learn much more by displaying the [bivariate](#) data in a graphical form that maintains the pairing. [\[link\]](#) shows a [scatter plot](#) of the paired ages. The x-axis represents the age of the husband and the y-axis the age of the wife.

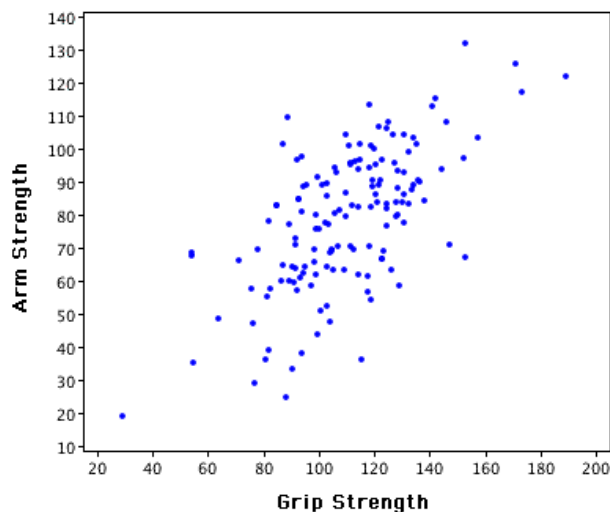


Scatterplot showing wife age as a function of husband age.

There are two important characteristics of the data revealed by [\[link\]](#). First, it is clear that there is a strong relationship between the husband's age and the wife's age: the older the husband, the older the wife. When one variable (y) increases with the second variable (x), we say that x and y have a [positive association](#). Conversely, when y decreases as x increases, we say that they have a [negative association](#).

Second, the points cluster along a straight line. When this occurs, the relationship is called a [linear relationship](#).

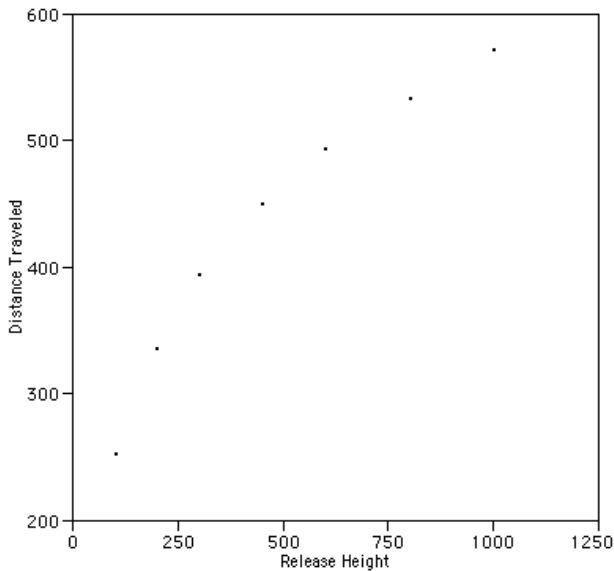
[\[link\]](#) shows a scatterplot of Arm Strength and Grip Strength from 149 individuals working in physically demanding jobs including electricians, construction and maintenance workers, and auto mechanics. Not surprisingly, the stronger someone's grip, the stronger their arm tends to be. There is therefore a positive association between these variables. Although the points cluster along a line, they are not clustered quite as closely as they are for the scatter plot of spousal age.



Scatter plot of Grip Strength and Arm Strength.

Not all scatter plots show linear relationships. [\[link\]](#) shows the results of an experiment conducted by Galileo on projectile motion. In the experiment, Galileo rolled balls down incline and measured how far they traveled as a function of the release height. It is clear from [\[link\]](#) that the relationship between "Release Height" and "Distance Traveled" is not described well by a straight line: If you drew a line connecting the lowest point and the highest point, all of the remaining points would be above the line. The data are better fit by a parabola.

- [D. Dickey and T. Arnold's description of the study including a movie](#)
- [Rice University's Galileo Project, section by S. Jennings](#)



Galileo's data showing a non-linear relationship.

Scatter plots that show linear relationships between variables can differ in several ways including the slope of the line about which they cluster and how tightly the points cluster about the line. A statistical measure of the strength of the relationship between variables that takes these factors into account is the subject of the next section.

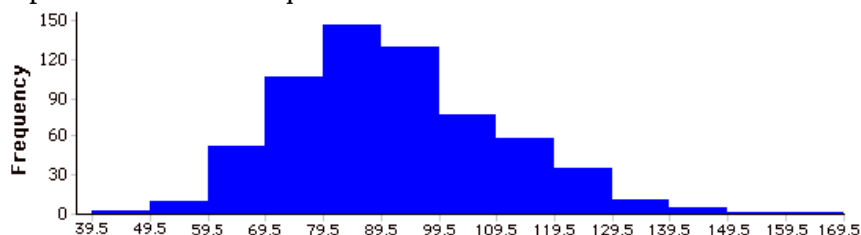
Glossary

Quantitative Variables

Variables that have are measured on a numeric or quantitative scale. Ordinal, interval and ratio scales are quantitative. A country's population, a person's shoe size, or a car's speed are all quantitative variables. Variables that are not quantitative are known as qualitative variables.

Histogram

A histogram is a graphical representation of a distribution. It partitions the variable on the x-axis into various contiguous class intervals of (usually) equal widths. The heights of the bars represent the class frequencies.



See also: [Sturgis's Rule](#)

Sturgis's Rule

One method of determining the number of classes for a [histogram](#), Sturgis's Rule is to take $1 + \log_2 N$ classes, rounded to the nearest integer.

Bivariate

Bivariate data is data for which there are two variables for each observation. As an example, the following bivariate data show the ages of husbands and wives of 10 married couples.

Husband	36	72	37	36	51	50	47	50	37	41
Wife	35	67	33	35	50	46	47	42	36	41

Scatter Plot

A scatter plot of two variables shows the values of one variable on the Y axis and the values of the other variable on the X axis. Scatter plots are well suited for revealing the relationship between two variables. The scatter plot shown in [\[link\]](#) illustrates data from one of Galileo's classic experiments in which he observed the distance traveled balls traveled after being dropped on a incline as a function of their release height.

Positive Association

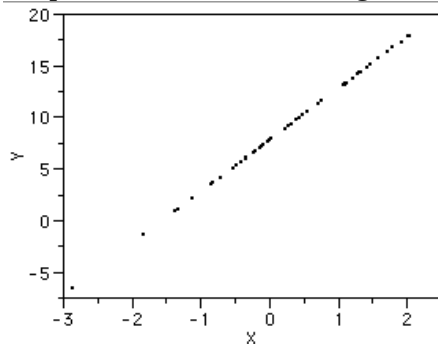
There is a positive association between variables X and Y if smaller values of X are associated with smaller values of Y and larger values of X are associated with larger values of Y .

Negative Association

There is a negative association between variables X and Y if smaller values of X are associated with larger values of Y and larger values of X are associated with smaller values of Y .

Linear Relationship

If the relationship between two variables is a perfect linear relationship, then a scatterplot of the points will fall on a straight line as shown in [\[link\]](#).



With real data, there is almost never a perfect linear relationship between two variables. The more the points tend to fall along a straight line the stronger the linear relationship. [\[link\]](#) shows two variables (husband's age and wife's age) that have a strong but not a perfect linear relationship.

Linear Regression and Correlation

This module provides an introduction of Linear Regression and Correlation as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret outliers.

Introduction

Professionals often want to know how two or more numeric variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is it and how strong is the relationship?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee. These are all examples in which regression can be used.

The type of data described in the examples is **bivariate** data - "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable (). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

Linear Regression and Correlation: Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. It has the form:

Equation:

$$y = a + bx$$

where a and b are constant numbers.

x is the independent variable, and y is the dependent variable.

Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example:

The following examples are linear equations.

Equation:

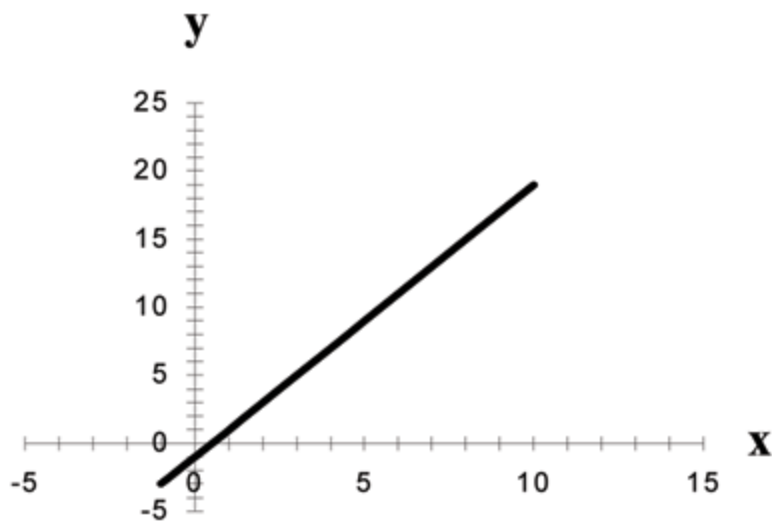
$$y = 3 + 2x$$

Equation:

$$y = -0.01 + 1.2x$$

The graph of a linear equation of the form $y = a + bx$ is a **straight line**. Any line that is not vertical can be described by this equation.

Example:



Graph of the equation $y = -1 + 2x$.

Linear equations of this form occur in applications of life sciences, social sciences, psychology, business, economics, physical sciences, mathematics, and other areas.

Example:

Aaron's Word Processing Service (AWPS) does word processing. Its rate is \$32 per hour plus a \$31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to do the word processing job.

Exercise:

Problem:

Find the equation that expresses the **total cost** in terms of the **number of hours** required to finish the word processing job.

Solution:

Let x = the number of hours it takes to get the job done.

Let y = the total cost to the customer.

The \$31.50 is a fixed cost. If it takes x hours to complete the job, then $(32)(x)$ is the cost of the word processing only. The total cost is:

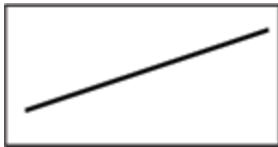
$$y = 31.50 + 32x$$

Linear Regression and Correlation: Slope and Y-Intercept of a Linear Equation

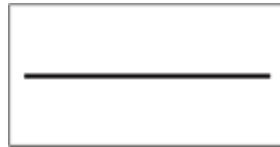
For the linear equation $y = a + bx$, b = slope and a = y-intercept.

From algebra recall that the slope is a number that describes the steepness of a line and the y-intercept is the y coordinate of the point a where the line crosses the y-axis.

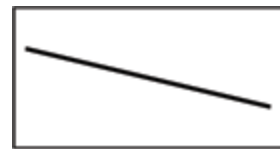
If $b > 0$, the line slopes upward to the right.



If $b = 0$, the line is horizontal.



If $b < 0$, the line slopes downward to the right.



Three possible graphs of $y = a + bx$.

Example:

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one time fee of \$25 plus \$15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

Exercise:

Problem:

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution:

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 ($a = 25$). At the start of the tutoring session, Svetlana charges a one-time fee of \$25 (this is when $x = 0$). The slope is 15 ($b = 15$). For each session, Svetlana earns \$15 for each hour she tutors.

Scatter Plots

This module provides an overview of Linear Regression and Correlation: Scatter Plots as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables x and y . The most common and easiest way is a **scatter plot**. The following example illustrates a scatter plot.

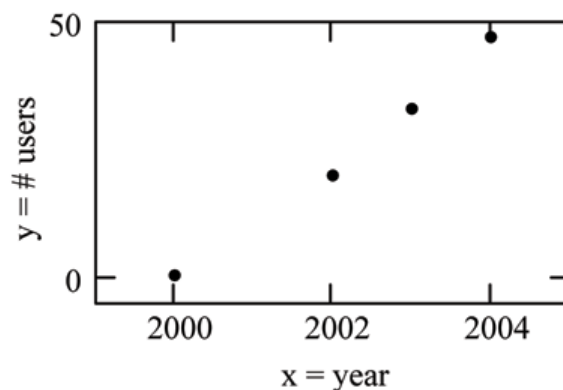
Example:

From an article in the *Wall Street Journal*: In Europe and Asia, m-commerce is popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. For the years 2000 through 2004, was there a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let x = the year and let y = the number of m-commerce users, in millions.

Table showing the number of m-commerce users (in millions) by year.

x (year)	y (# of users)
2000	0.5
2002	20.0

Scatter plot showing the number of m-commerce users (in millions) by year.



x (year)	y (# of users)
2003	33.0
2004	47.0

A scatter plot shows the **direction** and **strength** of a relationship between the variables. A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the strength of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function.

When you look at a scatterplot, you want to notice the **overall pattern** and any **deviations** from the pattern. The following scatterplot examples illustrate these concepts.

Positive Linear Pattern (Strong)



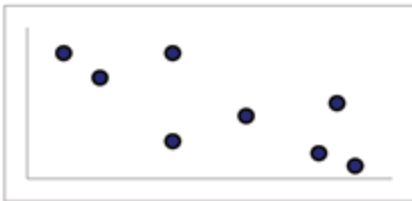
Linear Pattern w/ One Deviation



Positive Linear Pattern (Strong)



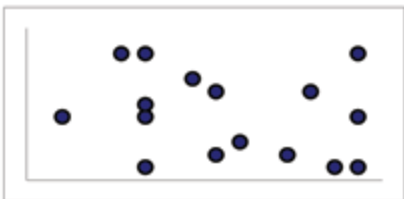
Negative Linear Pattern (Strong)



Exponential Growth Pattern



No Pattern



In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called **linear regression**. However, we only calculate a regression line if one of the variables helps to explain or predict

the other variable. If x is the independent variable and y the dependent variable, then we can use a regression line to predict y for a given value of x .

The Regression Equation

Linear Regression and Correlation: The Regression Equation is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean. Contributions from Roberta Bloom include instructions for finding and graphing the regression equation and scatterplot using the LinRegTTest on the TI-83,83+,84+ calculators.

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to "**fit**" a straight line. This is called a **Line of Best Fit or Least Squares Line**.

Optional Collaborative Classroom Activity

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x , is pinky finger length and the dependent variable, y , is height.

For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your lines so they cross the y-axis. Using the slopes and the y-intercepts, write your equation of "best fit". Do you think everyone will have the same equation? Why or why not?

Using your equation, what is the predicted height for a pinky length of 2.5 inches?

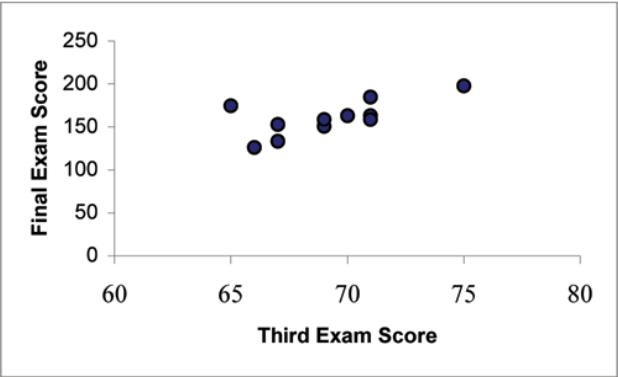
Example:

A random sample of 11 statistics students produced the following data where x is the third exam score, out of 80, and y is the final exam score, out of 200. Can you predict the final exam score of a random student if you know the third exam score?

Table showing the scores on the final exam based on scores from the third exam.

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

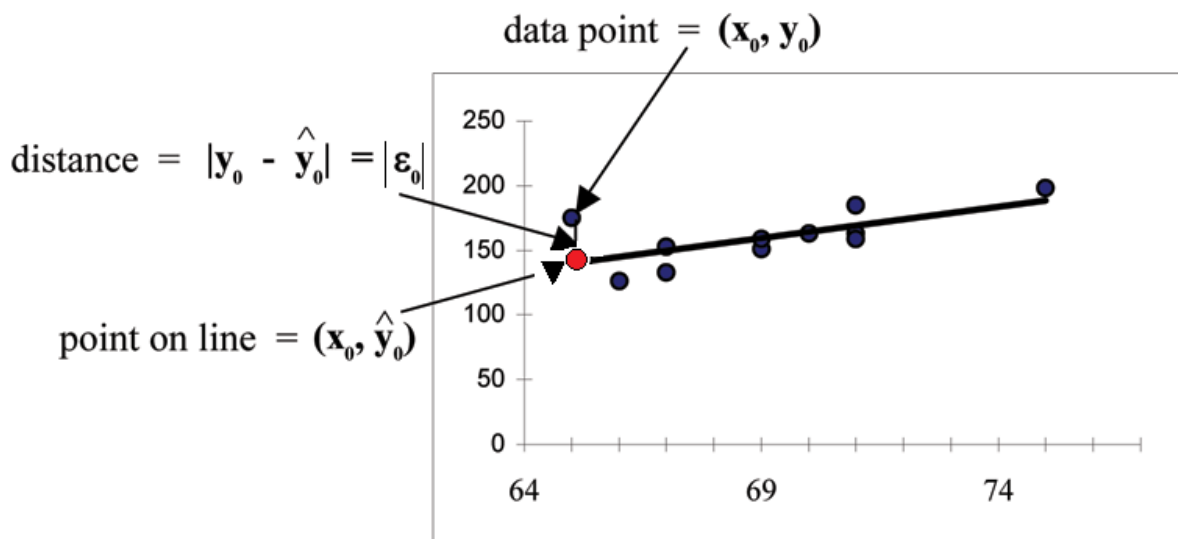
Scatter plot showing the scores on the final exam based on scores from the third exam.



The third exam score, x , is the independent variable and the final exam score, y , is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye", you would draw different lines. We can use what is called a **least-squares regression line** to obtain the best fit line.

Consider the following diagram. Each point of data is of the form (x, y) and each point of the line of best fit using least-squares linear regression has the form (x, \hat{y}) .

The \hat{y} is read "**y hat**" and is the **estimated value of y** . It is the value of y obtained using the regression line. It is not generally equal to y from data.



The term $y_0 - \hat{y}_0 = \epsilon_0$ is called the "**error**" or **residual**. It is not an error in the sense of a mistake. The **absolute value of a residual** measures the vertical distance between the actual value of y and the estimated value of y . In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for y . If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for y .

In the diagram above, $y_0 - \hat{y}_0 = \varepsilon_0$ is the residual for the point shown. Here the point lies above the line and the residual is positive.

ε = the Greek letter **epsilon**

For each data point, you can calculate the residuals or errors, $y_i - \hat{y}_i = \varepsilon_i$ for $i = 1, 2, 3, \dots, 11$.

Each $|\varepsilon|$ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ε values. If you square each ε and add, you get

$$\left(\varepsilon_1\right)^2 + \left(\varepsilon_2\right)^2 + \dots + \left(\varepsilon_{11}\right)^2 = \sum_{i=1}^{11} \varepsilon^2$$

This is called the **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the values of a and b that make the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

Equation:

$$\hat{y} = a + bx$$

where $a = \bar{y} - b \cdot \bar{x}$ and $b = \frac{\Sigma(x-\bar{x}) \cdot (y-\bar{y})}{\Sigma(x-\bar{x})^2}$.

\bar{x} and \bar{y} are the sample means of the x values and the y values, respectively. The best fit line always passes through the point (\bar{x}, \bar{y}) .

The slope b can be written as $b = r \cdot \frac{s_y}{s_x}$ where s_y = the standard deviation of the y values and s_x = the standard deviation of the x values. r is the correlation coefficient which is discussed in the next section.

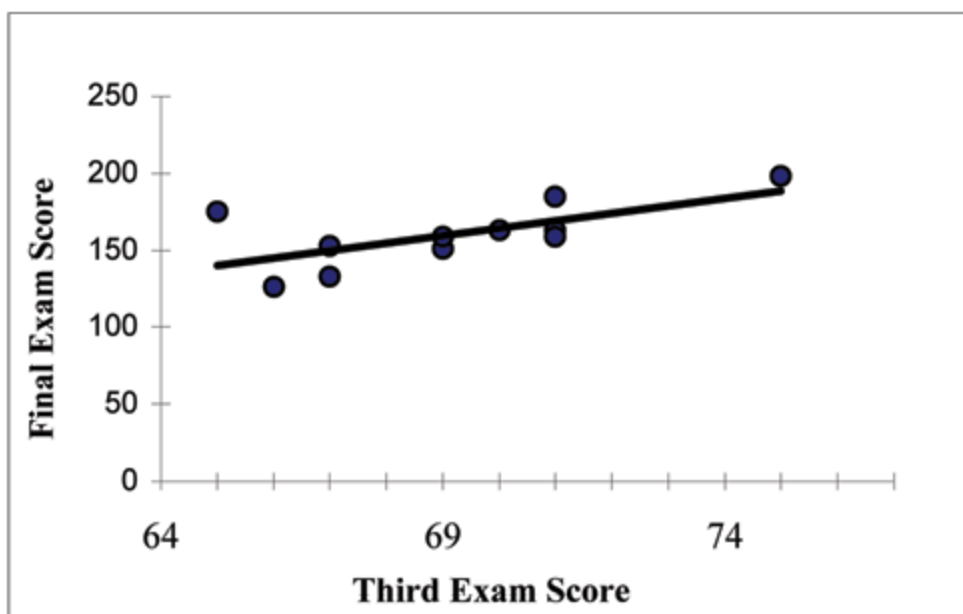
Least Squares Criteria for Best Fit

The process of fitting the best fit line is called **linear regression**. The idea behind finding the best fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the **least squares regression line**.

Note: Computer spreadsheets, statistical software, and many calculators can quickly calculate the best fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best fit line and create a scatterplot are shown at the end of this section.

THIRD EXAM vs FINAL EXAM EXAMPLE:

The graph of the line of best fit for the third exam/final exam example is shown below:



The least squares regression line (best fit line) for the third exam/final exam example has the equation:

Equation:

$$\hat{y} = -173.51 + 4.83x$$

Note:

- Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x -values in the sample data, **but not necessarily for x -values outside that domain.**
- You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam.
- You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x -values in the sample data, which are between 65 and 75.

UNDERSTANDING SLOPE

The slope of the line, b , describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

THIRD EXAM vs FINAL EXAM EXAMPLE

- Slope: The slope of the line is $b = 4.83$.

- Interpretation: For a one point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

Using the TI-83+ and TI-84+ Calculators

Using the Linear Regression T Test: LinRegTTest

In the STAT list editor, enter the X data in list L1 and the Y data in list L2, paired so that the corresponding (x,y) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.)

On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest as some calculators may also have a different item called LinRegTInt.)

On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1

On the next line, at the prompt β or ρ , highlight " $\neq 0$ " and press ENTER

Leave the line for "RegEq:" blank

Highlight Calculate and press ENTER.

LinRegTTest Input Screen and Output Screen

```

LinRegTTest
Xlist: L1
Ylist: L2
Freq: 1
 $\beta$  or  $\rho$  :  $\neq 0$  <0 >0
RegEQ:
Calculate
  
```

TI-83+ and TI-84+
calculators

```

LinRegTTest
y = a + bx
 $\beta \neq 0$  and  $\rho \neq 0$ 
t = 2.657560155
p = .0261501512
df = 9
↓ a = -173.513363
b = 4.827394209
s = 16.41237711
r2 = .4396931104
r = .663093591
  
```

The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items.

- The second line says $y=a+bx$. Scroll down to find the values $a=-173.513$, and $b=4.8273$; the equation of the best fit line is $\hat{y} = -173.51 + 4.83x$
- The two items at the bottom are $r^2 = .43969$ and $r=.663$. For now, just note where to find these values; we will discuss them in the next two sections.

Graphing the Scatterplot and Regression Line

We are assuming your X data is already entered in list L1 and your Y data is in list L2

Press 2nd STATPLOT ENTER to use Plot 1

On the input screen for PLOT 1, highlight **On** and press ENTER

For TYPE: highlight the very first icon which is the scatterplot and press ENTER

Indicate Xlist: L1 and Ylist: L2

For Mark: it does not matter which symbol you highlight.

Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data

To graph the best fit line, press the "Y=" key and type the equation $-173.5+4.83X$ into equation Y1. (The X key is immediately left of the STAT key). Press ZOOM 9 again to graph it.

Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax

****With contributions from Roberta Bloom**

Correlation Coefficient and Coefficient of Determination
 Linear Regression and Correlation: The Correlation Coefficient and Coefficient of Determination is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean with contributions from Roberta Bloom. The name has been changed from Correlation Coefficient.

The Correlation Coefficient r

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between x and y .

The **correlation coefficient, r** , developed by Karl Pearson in the early 1900s, is a numerical measure of the strength of association between the independent variable x and the dependent variable y .

The correlation coefficient is calculated as

Equation:

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

where n = the number of data points.

If you suspect a linear relationship between x and y , then r can measure how strong the linear relationship is.

What the VALUE of r tells us:

- The value of r is always between -1 and +1: $-1 \leq r \leq 1$.
- The size of the correlation r indicates the strength of the linear relationship between x and y . Values of r close to -1 or to +1 indicate a stronger linear relationship between x and y .
- If $r = 0$ there is absolutely no linear relationship between x and y (**no linear correlation**).

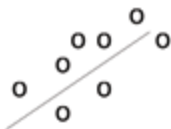
- If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

What the SIGN of r tells us

- A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease (**positive correlation**).
- A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase (**negative correlation**).
- The sign of r is the same as the sign of the slope, b , of the best fit line.

Note: Strong correlation does not suggest that x causes y or y causes x . We say "**correlation does not imply causation.**" For example, every person who learned math in the 17th century is dead. However, learning math does not necessarily cause death!

Positive Correlation



A scatter plot showing data with a positive correlation.

Negative Correlation



A scatter plot
showing data
with a
negative
correlation.

Zero Correlation



A scatter plot
showing data
with zero
correlation.
 $=0$

The formula for r looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate r . The correlation coefficient r is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions).

The Coefficient of Determination

r^2 is called the coefficient of determination. r^2 is the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. r^2 has an interpretation in the context of the data:

- r^2 , when expressed as a percent, represents the percent of variation in the dependent variable y that can be explained by variation in the independent variable x using the regression (best fit) line.
- $1-r^2$, when expressed as a percent, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

Consider the [third exam/final exam example](#) introduced in the previous section

- The line of best fit is: y
- The correlation coefficient is r
- The coefficient of determination is $r^2 = 0.4397$
- **Interpretation of r^2 in the context of this example:**
- Approximately 44% of the variation (0.4397 is approximately 0.44) in the final exam grades can be explained by the variation in the grades on the third exam, using the best fit regression line.
- Therefore approximately 56% of the variation ($1 - 0.44 = 0.56$) in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best fit regression line. (This is seen as the scattering of the points about the line.)

**With contributions from Roberta Bloom.

Glossary

Coefficient of Correlation

A measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable. The formula is:

Equation:

$$r = \frac{\frac{\sum n}{\sum x} - \frac{\sum x}{n}}{\frac{\sum x^2}{n} - \frac{(\sum x)^2}{n^2}} \quad \frac{\sum y}{n} - \frac{(\sum y)(\sum x)}{n^2}}{\frac{\sum y^2}{n} - \frac{(\sum y)^2}{n}}$$

where n is the number of data points. The coefficient cannot be more than 1 and less than -1. The closer the coefficient is to ± 1 , the stronger the evidence of a significant linear relationship between x and y .

Testing the Significance of the Correlation Coefficient
Linear Regression and Correlation: Testing the Significance of the Correlation Coefficient is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean. The title has been changed from Facts About the Correlation Coefficient for Linear Regression. Roberta Bloom has made major contributions to this module.

Testing the Significance of the Correlation Coefficient

The correlation coefficient, r , tells us about the strength of the linear relationship between x and y . However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient r and the sample size n , together.

We perform a hypothesis test of the "**significance of the correlation coefficient**" to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

The sample data is used to compute r , the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we only have sample data, we can not calculate the population correlation coefficient. The sample correlation coefficient, r , is our estimate of the unknown population correlation coefficient.

- The symbol for the population correlation coefficient is ρ , the Greek letter "rho".
- ρ = population correlation coefficient (unknown)
- r = sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to 0" or "significantly different from 0". We decide this based on the sample correlation coefficient r and the sample size n .

If the test concludes that the correlation coefficient is significantly different from 0, we say that the correlation coefficient is "significant".

- Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from 0."
- What the conclusion means: There is a significant linear relationship between x and y . We can use the regression line to model the linear relationship between x and y in the population.

If the test concludes that the correlation coefficient is not significantly different from 0 (it is close to 0), we say that correlation coefficient is "not significant".

- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from 0."
- What the conclusion means: There is not a significant linear relationship between x and y . Therefore we can NOT use the regression line to model a linear relationship between x and y in the population.

Note:

- If r is significant and the scatter plot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
- If r is not significant OR if the scatter plot does not show a linear trend, the line should not be used for prediction.
- If r is significant and if the scatter plot shows a linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

PERFORMING THE HYPOTHESIS TEST

SETTING UP THE HYPOTHESES:

- **Null Hypothesis:** $H_o: \rho = 0$
- **Alternate Hypothesis:** $H_a: \rho \neq 0$

What the hypotheses mean in words:

- **Null Hypothesis H_o :** The population correlation coefficient IS NOT significantly different from 0. There IS NOT a significant linear relationship(correlation) between x and y in the population.
- **Alternate Hypothesis H_a :** The population correlation coefficient IS significantly DIFFERENT FROM 0. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between x and y in the population.

DRAWING A CONCLUSION:

- There are two methods to make the decision. Both methods are equivalent and give the same result.
- **Method 1: Using the p-value**
- **Method 2: Using a table of critical values**
- In this chapter of this textbook, we will always use a significance level of 5%, $\alpha = 0.05$
- Note: Using the p-value method, you could choose any appropriate significance level you want; you are not limited to using $\alpha = 0.05$. But the table of critical values provided in this textbook assumes that we are using a significance level of 5%, $\alpha = 0.05$. (If we wanted to use a different significance level than 5% with the critical value method, we would need different tables of critical values that are not provided in this textbook.)

METHOD 1: Using a p-value to make a decision

- The linear regression t -test LinRegTTEST on the TI-83+ or TI-84+ calculators calculates the p-value.
- On the LinRegTTEST input screen, on the line prompt for β or ρ , highlight " $\neq 0$ "

- The output screen shows the p-value on the line that reads "p =".
- (Most computer statistical software can calculate the p-value.)

If the p-value is less than the significance level ($\alpha = 0.05$):

- Decision: REJECT the null hypothesis.
- Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from 0."

If the p-value is NOT less than the significance level ($\alpha = 0.05$)

- Decision: DO NOT REJECT the null hypothesis.
- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is NOT significantly different from 0."

Calculation Notes:

- You will use technology to calculate the p-value. The following describe the calculations to compute the test statistics and the p-value:
- The p-value is calculated using a t -distribution with $n-2$ degrees of freedom.
- The formula for the test statistic is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. The value of the test statistic, t , is shown in the computer or calculator output along with the p-value. The test statistic t has the same sign as the correlation coefficient r .
- The p-value is the combined area in both tails.
- An alternative way to calculate the p-value (**p**) given by LinRegTTest is the command `2*tcdf(abs(t),10^99,n-2)` in 2nd DISTR.

THIRD EXAM vs FINAL EXAM EXAMPLE: p value method

- Consider the [third exam/final exam example](#).
- The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are $n = 11$ data points.

- Can the regression line be used for prediction? **Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?**
- $H_o: \rho = 0$
- $H_a: \rho \neq 0$
- $\alpha = 0.05$
- The p-value is 0.026 (from LinRegTTest on your calculator or from computer software)
- The p-value, 0.026, is less than the significance level of $\alpha = 0.05$
- Decision: Reject the Null Hypothesis H_o
- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from 0.
- **Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.**

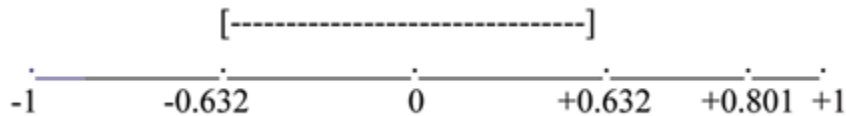
METHOD 2: Using a table of Critical Values to make a decision

The [95% Critical Values of the Sample Correlation Coefficient Table](#) at the end of this chapter (before the [Summary](#)) may be used to give you a good idea of whether the computed value of r is **significant or not**.

Compare r to the appropriate critical value in the table. If r is not between the positive and negative critical values, then the correlation coefficient is significant. If r is significant, then you may want to use the line for prediction.

Example:

Suppose you computed $r = 0.801$ using $n = 10$ data points.
 $df = n - 2 = 10 - 2 = 8$. The critical values associated with $df = 8$ are -0.632 and + 0.632. If $r <$ negative critical value or $r >$ positive critical value, then r is significant. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you.



r is not significant between -0.632 and $+0.632$.
 $r = 0.801 > +0.632$. Therefore, r is significant.

Example:

Suppose you computed $r = -0.624$ with 14 data points.
 $df = 14 - 2 = 12$. The critical values are -0.532 and 0.532 . Since $-0.624 < -0.532$, r is significant and the line may be used for prediction



$r = -0.624 < -0.532$. Therefore, r is significant.

Example:

Suppose you computed $r = 0.776$ and $n = 6$. $df = 6 - 2 = 4$. The critical values are -0.811 and 0.811 . Since $-0.811 < 0.776 < 0.811$, r is not significant and the line should not be used for prediction.



$-0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.

THIRD EXAM vs FINAL EXAM EXAMPLE: critical value method

- Consider the [third exam/final exam example](#).
- The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$ and there are $n = 11$ data points.
- Can the regression line be used for prediction? **Given a third exam score (x value), can we use the line to predict the final exam score (predicted y value)?**
- $H_o: \rho = 0$
- $H_a: \rho \neq 0$
- $\alpha = 0.05$
- Use the "95% Critical Value" table for r with $df = n - 2 = 11 - 2 = 9$
- The critical values are -0.602 and $+0.602$
- Since $0.6631 > 0.602$, r is significant.
- Decision: Reject H_o :
- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from 0.
- **Because r is significant and the scatter plot shows a linear trend, the regression line can be used to predict final exam scores.**

Example:

Additional Practice Examples using Critical Values

Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine if r is significant and the line of best fit associated with each r can be used to predict a y value. If it helps, draw a number line.

1. $r = -0.567$ and the sample size, n , is 19. The $df = n - 2 = 17$. The critical value is -0.456 . $-0.567 < -0.456$ so r is significant.
2. $r = 0.708$ and the sample size, n , is 9. The $df = n - 2 = 7$. The critical value is 0.666 . $0.708 > 0.666$ so r is significant.
3. $r = 0.134$ and the sample size, n , is 14. The $df = 14 - 2 = 12$. The critical value is 0.532 . 0.134 is between -0.532 and 0.532 so r is not

significant.

4. $r = 0$ and the sample size, n , is 5. No matter what the dfs are, $r = 0$ is between the two critical values so r is not significant.

Assumptions in Testing the Significance of the Correlation Coefficient

Testing the significance of the correlation coefficient requires that certain assumptions about the data are satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population. We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between x and y in the sample data provides strong enough evidence so that we can conclude that there is a linear relationship between x and y in the population.

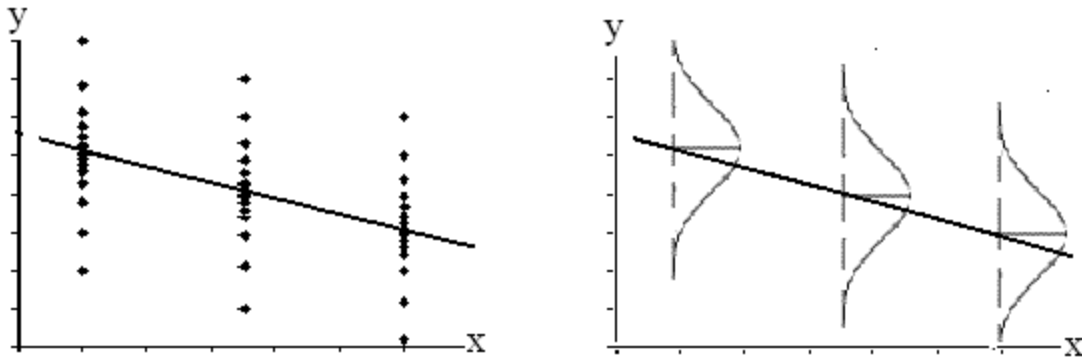
The regression line equation that we calculate from the sample data gives the best fit line for our particular sample. We want to use this best fit line for the sample as an estimate of the best fit line for the population.

Examining the scatterplot and testing the significance of the correlation coefficient helps us determine if it is appropriate to do this.

The assumptions underlying the test of significance are:

- There is a linear relationship in the population that models the average value of y for varying values of x . In other words, the expected value of y for each particular value lies on a straight line in the population. (We do not know the equation for the line for the population. Our regression line from the sample is our best estimate of this line in the population.)
- The y values for any particular x value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away. Assumption (1) above implies that these normal distributions are centered on the line: the means of these normal distributions of y values lie on the line.

- The standard deviations of the population y values about the line are equal for each value of x . In other words, each of these normal distributions of y values has the same shape and spread about the line.
- The residual errors are mutually independent (no pattern).



The y values for each x value are normally distributed about the line with the same standard deviation. For each x value, the mean of the y values lies on the regression line. More y values lie near the line than are scattered further away from the line.

**With contributions from Roberta Bloom

Prediction

Linear Regression and Correlation: Prediction is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean with contributions from Roberta Bloom.

Recall the [third exam/final exam example](#).

We examined the scatterplot and showed that the correlation coefficient is significant. We found the equation of the best fit line for the final exam grade as a function of the grade on the third exam. We can now use the least squares regression line for prediction.

Suppose you want to estimate, or predict, the final exam score of statistics students who received 73 on the third exam. The exam scores (***x*-values**) range from 65 to 75. **Since 73 is between the *x*-values 65 and 75,** substitute $x = 73$ into the equation. Then:

Equation:

$$\hat{y} = -173.51 + 4.83(73) = 179.08$$

We predict that statistic students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

Example:

Recall the [third exam/final exam example](#).

Exercise:

Problem:

What would you predict the final exam score to be for a student who scored a 66 on the third exam?

Solution:

145.27

Exercise:

Problem:

What would you predict the final exam score to be for a student who scored a 90 on the third exam?

Solution:

The x values in the data are between 65 and 75. 90 is outside of the domain of the observed x values in the data (independent variable), so you cannot reliably predict the final exam score for this student. (Even though it is possible to enter x into the equation and calculate a y value, you should not do so!)

To really understand how unreliable the prediction can be outside of the observed x values in the data, make the substitution $x = 90$ into the equation.

$$\hat{y} = -173.51 + 4.83(90) = 261.19$$

The final exam score is predicted to be 261.19. The largest the final exam score can be is 200.

Note: The process of predicting inside of the observed x values in the data is called **interpolation**. The process of predicting outside of the observed x values in the data is called **extrapolation**.

Outliers

Linear Regression and Correlation: Outliers is a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean. The module has been modified to include a graphical method for identifying outliers contributed by Roberta Bloom.

In some data sets, there are values (**observed data points**) called [outliers](#). **Outliers are observed data points that are far from the least squares line.** They have large "errors", where the "error" or residual is the vertical distance from the line to the point.

Outliers need to be examined closely. Sometimes, for some reason or another, they should not be included in the analysis of the data. It is possible that an outlier is a result of erroneous data. Other times, an outlier may hold valuable information about the population under study and should remain included in the data. The key is to carefully examine what causes a data point to be an outlier.

Besides outliers, a sample may contain one or a few points that are called **influential points**. Influential points are observed data points that are far from the other observed data points in the horizontal direction. These points may have a big effect on the slope of the regression line. To begin to identify an influential point, you can remove it from the data set and see if the slope of the regression line is changed significantly.

Computers and many calculators can be used to identify outliers from the data. Computer output for regression analysis will often identify both outliers and influential points so that you can examine them.

Identifying Outliers

We could guess at outliers by looking at a graph of the scatterplot and best fit line. However we would like some guideline as to how far away a point needs to be in order to be considered an outlier. **As a rough rule of thumb, we can flag any point that is located further than two standard deviations above or below the best fit line as an outlier.** The standard deviation used is the standard deviation of the residuals or errors.

We can do this visually in the scatterplot by drawing an extra pair of lines that are two standard deviations above and below the best fit line. Any data points that are outside this extra pair of lines are flagged as potential outliers. Or we can do this numerically by calculating each residual and comparing it to twice the standard deviation. On the TI-83, 83+, or 84+, the graphical approach is easier. The graphical procedure is shown first, followed by the numerical calculations. You would generally only need to use one of these methods.

Example:

Exercise:

Problem:

In the [third exam/final exam example](#), you can determine if there is an outlier or not. If there is an outlier, as an exercise, delete it and fit the remaining data to a new line. For this example, the new line ought to fit the remaining data better. This means the **SSE** should be smaller and the correlation coefficient ought to be closer to 1 or -1.

Solution:

Graphical Identification of Outliers

With the TI-83,83+,84+ graphing calculators, it is easy to identify the outlier graphically and visually. If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance was equal to $2s$ or farther, then we would consider the data point to be "too far" from the line of best fit. We need to find and graph the lines that are two standard deviations below and above the regression line. Any points that are outside these two lines are outliers. We will call these lines Y_2 and Y_3 :

As we did with the equation of the regression line and the correlation coefficient, we will use technology to calculate this standard deviation for us. Using the **LinRegTTest** with this data, scroll down through the output screens to find **$s=16.412$**

Line $Y_2 = -173.5 + 4.83x - 2(16.4)$ and line $Y_3 = -173.5 + 4.83x + 2(16.4)$

where $\hat{y} = -173.5 + 4.83x$ is the line of best fit. Y_2 and Y_3 have the same slope as the line of best fit.

Graph the scatterplot with the best fit line in equation Y_1 , then enter the two extra lines as Y_2 and Y_3 in the "Y=" equation editor and press ZOOM 9. You will find that the only data point that is not between lines Y_2 and Y_3 is the point $x=65$, $y=175$. On the calculator screen it is just barely outside these lines. The outlier is the student who had a grade of 65 on the third exam and 175 on the final exam; this point is further than 2 standard deviations away from the best fit line.

Sometimes a point is so close to the lines used to flag outliers on the graph that it is difficult to tell if the point is between or outside the lines. On a computer, enlarging the graph may help; on a small calculator screen, zooming in may make the graph clearer. Note that when the graph does not give a clear enough picture, you can use the numerical comparisons to identify outliers.

[missing_resource: linrgoutlier.gif]

Numerical Identification of Outliers

In the table below, the first two columns are the third exam and final exam data. The third column shows the predicted \hat{y} values calculated from the line of best fit: $\hat{y} = -173.5 + 4.83x$. The residuals, or errors, have been calculated in the fourth column of the table:

observed y value – predicted y value = $y - \hat{y}$.

s is the standard deviation of all the $y - \hat{y} = \varepsilon$ values where n = the total number of data points. If each residual is calculated and squared, and the results are added, we get the SSE. The standard deviation of the residuals is calculated from the SSE as:

$$s = \sqrt{\frac{\text{SSE}}{n-2}}$$

Rather than calculate the value of s ourselves, we can find s using the computer or calculator. For this example, the calculator function

LinRegTTest found $s = 16.4$ as the standard deviation of the residuals
 35 -17 16 -6 -19 9 3 -1 -10 -9 -1 .

x	y	\hat{y}	$y - \hat{y}$
65	175	140	$175 - 140 = 35$
67	133	150	$133 - 150 = -17$
71	185	169	$185 - 169 = 16$
71	163	169	$163 - 169 = -6$
66	126	145	$126 - 145 = -19$
75	198	189	$198 - 189 = 9$
67	153	150	$153 - 150 = 3$
70	163	164	$163 - 164 = -1$
71	159	169	$159 - 169 = -10$
69	151	160	$151 - 160 = -9$
69	159	160	$159 - 160 = -1$

We are looking for all data points for which the residual is greater than $2s=2(16.4)=32.8$ or less than -32.8 . Compare these values to the residuals in column 4 of the table. The only such data point is the student who had a grade of 65 on the third exam and 175 on the final exam; the residual for this student is 35.

How does the outlier affect the best fit line?

Numerically and graphically, we have identified the point (65,175) as an outlier. We should re-examine the data for this point to see if there are any problems with the data. If there is an error we should fix the error if possible, or delete the data. If the data is correct, we would leave it in the data set. **For this problem, we will suppose that we examined the data and found that this outlier data was an error. Therefore we will continue on and delete the outlier, so that we can explore how it affects the results, as a learning experience.**

Compute a new best-fit line and correlation coefficient using the 10 remaining points:

On the TI-83, TI-83+, TI-84+ calculators, delete the outlier from L1 and L2. Using the LinRegTTest, the new line of best fit and the correlation coefficient are:

$$\hat{y} = -355.19 + 7.39x \text{ and } r = 0.9121$$

The new line with $r = 0.9121$ is a stronger correlation than the original ($r=0.6631$) because $r = 0.9121$ is closer to 1. This means that the new line is a better fit to the 10 remaining data values. The line can better predict the final exam score given the third exam score.

Numerical Identification of Outliers: Calculating s and Finding Outliers Manually

If you do not have the function LinRegTTest, then you can calculate the outlier in the first example by doing the following.

First, **square each** $y - \hat{y}$ (See the TABLE above):

The squares are 35^2 17^2 16^2 6^2 19^2 9^2 3^2 1^2 10^2 9^2 1^2

Then, add (sum) all the $y - \hat{y}$ **squared terms** using the formula

$$\sum_{i=1}^{11} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{11} \varepsilon_i^2 \quad (\text{Recall that } y_i - \hat{y}_i = \varepsilon_i.)$$

$$= 35^2 + 17^2 + 16^2 + 6^2 + 19^2 + 9^2 + 3^2 + 1^2 + 10^2 + 9^2 + 1^2$$

$= 2440 = \text{SSE}$. The result, **SSE** is the Sum of Squared Errors.

Next, calculate s , the standard deviation of all the $y - \hat{y} = \varepsilon$ values where $n =$ the total number of data points.

The calculation is $s = \sqrt{\frac{\text{SSE}}{n-2}}$

For the third exam/final exam problem, $s = \sqrt{\frac{2440}{11-2}} = 16.47$

Next, multiply s by 1.9:

$$(1.9) \cdot (16.47) = 31.29$$

31.29 is almost 2 standard deviations away from the mean of the $y - \hat{y}$ values.

If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance is at least $1.9s$, then we would consider the data point to be "too far" from the line of best fit. We call that point a **potential outlier**.

For the example, if any of the $y - \hat{y}$ values are **at least** 31.29, the corresponding (x, y) data point is a potential outlier.

For the third exam/final exam problem, all the $y - \hat{y}$'s are less than 31.29 except for the first one which is 35.

$$35 > 31.29 \quad \text{That is, } y - \hat{y} \geq 1.9 \cdot s$$

The point which corresponds to $y - \hat{y} = 35$ is $(65, 175)$. **Therefore, the data point $(65, 175)$ is a potential outlier.** For this example, we will delete it. (Remember, we do not always delete an outlier.)

The next step is to compute a new best-fit line using the 10 remaining points. The new line of best fit and the correlation coefficient are:

$$\hat{y} = -355.19 + 7.39x \text{ and } r = 0.9121$$

Example:**Exercise:****Problem:**

Using this new line of best fit (based on the remaining 10 data points), what would a student who receives a 73 on the third exam expect to receive on the final exam? Is this the same as the prediction made using the original line?

Solution:

Using the new line of best fit, $\hat{y} = -355.19 + 7.39(73) = 184.28$. A student who scored 73 points on the third exam would expect to earn 184 points on the final exam.

The original line predicted $\hat{y} = -173.51 + 4.83(73) = 179.08$ so the prediction using the new line with the outlier eliminated differs from the original prediction.

Example:

(From The Consumer Price Indexes Web site) The Consumer Price Index (CPI) measures the average change over time in the prices paid by urban consumers for consumer goods and services. The CPI affects nearly all Americans because of the many ways it is used. One of its biggest uses is as a measure of inflation. By providing information about price changes in the Nation's economy to government, business, and labor, the CPI helps them to make economic decisions. The President, Congress, and the

Federal Reserve Board use the CPI's trends to formulate monetary and fiscal policies. In the following table, x is the year and y is the CPI.

x	y
1915	10.1
1926	17.7
1935	13.7
1940	14.7
1947	24.1
1952	26.5
1964	31.0
1969	36.7
1975	49.3
1979	72.6
1980	82.4
1986	109.6
1991	130.7
1999	166.6

Data:

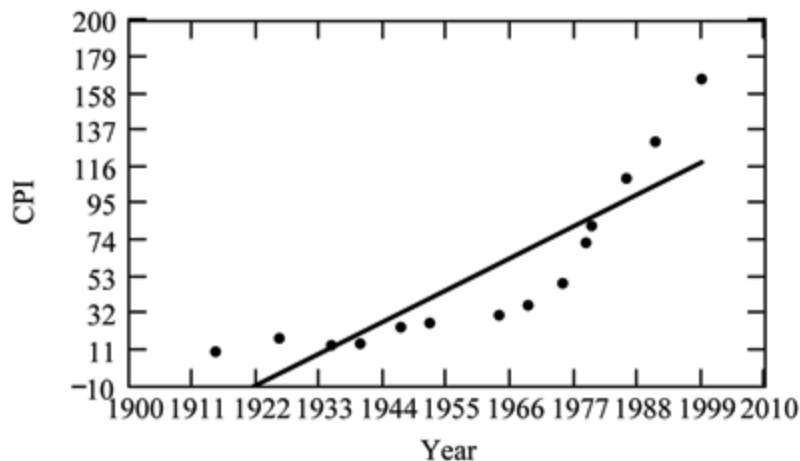
Exercise:

Problem:

- Make a scatterplot of the data.
- Calculate the least squares line. Write the equation in the form $\hat{y} = a + bx$.
- Draw the line on the scatterplot.
- Find the correlation coefficient. Is it significant?
- What is the average CPI for the year 1990?

Solution:

- Scatter plot and line of best fit.
- $\hat{y} = -3204 + 1.662x$ is the equation of the line of best fit.
- $r = 0.8694$
- The number of data points is $n = 14$. Use the 95% Critical Values of the Sample Correlation Coefficient table at the end of Chapter 12. $n - 2 = 12$. The corresponding critical value is 0.532. Since $0.8694 > 0.532$, r is significant.
- $\hat{y} = -3204 + 1.662(1990) = 103.4$ CPI
- Using the calculator LinRegTTest, we find that $s = 25.4$; graphing the lines $Y2 = -3204 + 1.662X - 2(25.4)$ and $Y3 = -3204 + 1.662X + 2(25.4)$ shows that no data values are outside those lines, identifying no outliers. (Note that the year 1999 was very close to the upper line, but still inside it.)



Note: In the example, notice the pattern of the points compared to the line. Although the correlation coefficient is significant, the pattern in the scatterplot indicates that a curve would be a more appropriate model to use than a line. In this example, a statistician should prefer to use other methods to fit a curve to this data, rather than model the data with the line we found. In addition to doing the calculations, it is always important to look at the scatterplot when deciding whether a linear model is appropriate.

If you are interested in seeing more years of data, visit the Bureau of Labor Statistics CPI website <ftp://ftp.bls.gov/pub/special.requests/cpi/cpiai.txt> ; our data is taken from the column entitled "Annual Avg." (third column from the right). For example you could add more current years of data. Try adding the more recent years 2004 : CPI=188.9, 2008 : CPI=215.3 and 2011: CPI=224.9. See how it affects the model. (Check: $\hat{y} = -4436 + 2.295x$. $r = 0.9018$. Is r significant? Is the fit better with the addition of the new points?)

**With contributions from Roberta Bloom

Glossary

Outlier

An observation that does not fit the rest of the data.

95% Critical Values of the Sample Correlation Coefficient Table
This module provides an overview of Linear Regression and Correlation:
95% Critical Values of the Sample Correlation Coefficient Table as a part of
Collaborative Statistics collection (col10522) by Barbara Illowsky and
Susan Dean.

Degrees of Freedom:	Critical Values: (and)
1	0.997
2	0.950
3	0.878
4	0.811
5	0.754
6	0.707
7	0.666
8	0.632
9	0.602
10	0.576
11	0.555
12	0.532

Degrees of Freedom:	Critical Values: (and)
13	0.514
14	0.497
15	0.482
16	0.468
17	0.456
18	0.444
19	0.433
20	0.423
21	0.413
22	0.404
23	0.396
24	0.388
25	0.381
26	0.374
27	0.367
28	0.361
29	0.355

Degrees of Freedom:	Critical Values: (and)
30	0.349
40	0.304
50	0.273
60	0.250
70	0.232
80	0.217
90	0.205
100	0.195

Linear Regression and Correlation: Summary

Bivariate Data: Each data point has two values. The form is (x, y) .

Line of Best Fit or Least Squares Line (LSL): $\hat{y} = a + bx$

x = independent variable; y = dependent variable

Residual: Actual y value $-$ predicted y value $= y - \hat{y}$

Correlation Coefficient r :

1. Used to determine whether a line of best fit is good for prediction.
2. Between -1 and 1 inclusive. The closer r is to 1 or -1, the closer the original points are to a straight line.
3. If r is negative, the slope is negative. If r is positive, the slope is positive.
4. If $r = 0$, then the line is horizontal.

Sum of Squared Errors (SSE): The smaller the SSE, the better the original set of points fits the line of best fit.

Outlier: A point that does not seem to fit the rest of the data.

The Coefficient of Determination: Family Wealth and Student Achievement Scores



Note: This module has been peer-reviewed, accepted, and sanctioned by the National Council of Professors of Educational Administration (NCPEA) as a scholarly contribution to the knowledge base in educational administration.

The “coefficient of determination” is equal to the correlation squared. When multiplied by 100, the coefficient of determination becomes the percentage of the variance that is associated with, determined by, or accounted for by the variance.

For example, if the correlation between two variables (X and Y) is .5, and if a causal relationship between the two variables can be established, then the percentage of the variance in Y that is accounted for by the variance in X is 25, or one-fourth.

Application:

The “No Child Left Behind” Act requires an annual review of each school served. If adequate yearly progress by a school is not made for two consecutive years, the school is designated for “school improvement.” A school that continues to fail to achieve adequate yearly progress for two years after being designated for school improvement must be identified by the local education agency for “corrective action.” If after being designated for “corrective action” a school fails to make adequate yearly progress, the school is to be designated for “restructuring.”

Teaching or administering a school designated for “school improvement,” “corrective action,” or “restructuring” is professionally embarrassing. However, since both James Coleman (1966) and Christopher Jenks (1972) found that there is often a strong correlation between family wealth and student standardized achievement test scores, punishing low-achieving schools in economically poor neighborhoods is questionable at best.

Coleman collected data on 600,000 children in all fifty states. He noticed that there were large differences in school quality, and believed that this was because schools in the affluent suburbs were well financed, whereas schools in the inner cities were deteriorating. The Civil Rights Act of 1964 ordered the Commissioner of Education to investigate, and Coleman was asked to head that investigation. He predicted that it was the difference in the quality of schools that accounted for the difference in the academic achievement of the poor and minorities.

To his surprise, he found that non-school factors, particularly family background, accounted for the difference:

One implication stands out above all: That schools bring little influence to bear on a child’s achievement that is independent of his background and general social context; and that this very lack of an independent effect means that the inequalities imposed on children by their home, neighborhood, and peer environment are carried along to become the inequalities with which they confront adult life at the end of school. (Coleman, 1966)

A subsequent large three-year study by Christopher Jenks confirmed Coleman’s findings. (Jenks, 1972)

Additional Evidence:

Given the findings of Coleman and Jenks, here is additional evidence, utilizing data from high schools in Kern County (California):

High School	2006 California Academic Performance Index	2006 Percentage of Students Qualifying for Free and Reduced Meals
Stockdale High	770	12.2
Burroughs High	746	24.5
Desert High	736	10.7
Liberty High	710	9.1
Tehachapi High	708	25.9
Frazier Mountain High	699	37.3
Kern Valley High	698	46.2
Centennial High	694	22.6
Bakersfield High	672	43.1
Boron High	668	39.4
Delano	667	79.3

High		
Ridgeview High	664	39.0
Chavez High	663	76.0
North High	658	45.6
Rosamond High	657	47.4
Mojave High	653	54.8
Shafter High	650	68.8
Taft High	650	50.3
Highland High	636	51.5
West High	634	53.5
McFarland High	617	79.3
Golden Valley High	614	82.2
Wasco High	600	73.9
Arvin High	600	82.2

Foothill High	596	62.3
East Bakersfield High	592	55.7

The correlation between these two variables is $-.812$ (Notice that this inverse relationship is not perfect, in which case the correlation would have been -1.00 , but very close.) Squaring this correlation coefficient allows one to compute the coefficient of determination. In this case the coefficient of determination is $.6592$. Multiplying by 100 (and rounding), this suggests that about 66% of the variability in the achievement test scores is strongly related to family wealth. High scoring schools have a low percentage of students qualifying for free and reduced price meals, whereas low scoring schools have a high percentage of students qualifying for free and reduced price meals.

This implies that family wealth—a variable not controlled by educators—is having a large impact on student achievement test scores. It suggests that improving student academic achievement in impoverished areas is, and will be, very difficult.

References

Coleman, James et al. (1966). *Equality of Educational Opportunity*. Washington, D.C.: U.S. Government Printing Office, 235.

Jenks, Christopher (1972). *Inequality*. New York: Harper & Row.

Hypothesis Testing: Single Mean and Single Proportion

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Differentiate between Type I and Type II Errors
- Describe hypothesis testing in general and in practice
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation known.
- Conduct and interpret hypothesis tests for a single population mean, population standard deviation unknown.
- Conduct and interpret hypothesis tests for a single population proportion.

Introduction

One job of a statistician is to make statistical inferences about populations based on samples taken from the population. **Confidence intervals** are one way to estimate a population parameter. Another way to make a statistical inference is to make a decision about a parameter. For instance, a car dealer advertises that its new small truck gets 35 miles per gallon, on the average. A tutoring service claims that its method of tutoring helps 90% of its students get an A or a B. A company says that women managers in their company earn an average of \$60,000 per year.

A statistician will make a decision about these claims. This process is called **"hypothesis testing."** A hypothesis test involves collecting data from a sample and evaluating the data. Then, the statistician makes a decision as to whether or not there is sufficient evidence based upon analyses of the data, to reject the null hypothesis.

In this chapter, you will conduct hypothesis tests on single means and single proportions. You will also learn about the errors associated with these tests.

Hypothesis testing consists of two contradictory hypotheses or statements, a decision based on the data, and a conclusion. To perform a hypothesis test, a

statistician will:

1. Set up two contradictory hypotheses.
2. Collect sample data (in homework problems, the data or summary statistics will be given to you).
3. Determine the correct distribution to perform the hypothesis test.
4. Analyze sample data by performing the calculations that ultimately will allow you to reject or fail to reject the null hypothesis.
5. Make a decision and write a meaningful conclusion.

Note: To do the hypothesis test homework problems for this chapter and later chapters, make copies of the appropriate special solution sheets. See the Table of Contents topic "Solution Sheets".

Glossary

Confidence Interval (CI)

An interval estimate for an unknown population parameter. This depends on:

- The desired confidence level.
- Information that is known about the distribution (for example, known standard deviation).
- The sample and its size.

Hypothesis Testing

Based on sample evidence, a procedure to determine whether the hypothesis stated is a reasonable statement and cannot be rejected, or is unreasonable and should be rejected.

Null and Alternate Hypotheses

The actual test begins by considering two [hypotheses](#). They are called the **null hypothesis** and the **alternate hypothesis**. These hypotheses contain opposing viewpoints.

H_o : **The null hypothesis:** It is a statement about the population that will be assumed to be true unless it can be shown to be incorrect beyond a reasonable doubt.

H_a : **The alternate hypothesis:** It is a claim about the population that is contradictory to H_o and what we conclude when we reject H_o .

Example:

H_o : No more than 30% of the registered voters in Santa Clara County voted in the primary election.

H_a : More than 30% of the registered voters in Santa Clara County voted in the primary election.

Example:

We want to test whether the mean grade point average in American colleges is different from 2.0 (out of 4.0).

$H_o: \mu$ $H_a: \mu$

Example:

We want to test if college students take less than five years to graduate from college, on the average.

$H_o: \mu$ $H_a: \mu$

Example:

In an issue of **U. S. News and World Report**, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams and a third pass. The same article stated that 6.6% of U. S. students take advanced placement exams and 4.4 % pass. Test if the percentage of U. S. students who take advanced placement exams is more than 6.6%.

$H_o: p$

$H_a: p$

Since the null and alternate hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make a **decision**. There are two options for a decision. They are "reject H_o " if the sample information favors the alternate hypothesis or "do not reject H_o " or "fail to reject H_o " if the sample information is insufficient to reject the null hypothesis.

Mathematical Symbols Used in H_o and H_a :

H_o	H_a
equal ()	not equal () or greater than () or less than ()
greater than or equal to ()	less than ()
less than or equal to ()	more than ()

Note: H_o always has a symbol with an equal in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test. However, be aware that many researchers (including one of the co-authors in research work) use H_o in the Null Hypothesis, even with H_a or H_1 as the symbol in the Alternate Hypothesis. This practice is acceptable because we only make the decision to reject or not reject the Null Hypothesis.

Optional Collaborative Classroom Activity

Bring to class a newspaper, some news magazines, and some Internet articles . In groups, find articles from which your group can write a null and alternate hypotheses. Discuss your hypotheses with the rest of the class.

Glossary

Hypothesis

A statement about the value of a population parameter. In case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_o) and the contradictory statement is called the alternate hypothesis (notation H_a).

Outcomes and the Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_o and the decision to reject or not. The outcomes are summarized in the following table:

ACTION	H_o IS ACTUALLY	...
	True	False
Do not reject H_o	Correct Outcome	Type II error
Reject H_o	Type I Error	Correct Outcome

The four possible outcomes in the table are:

- The decision is to **not reject H_o** when, in fact, H_o is **true (correct decision)**.
- The decision is to **reject H_o** when, in fact, H_o is **true** (incorrect decision known as a **Type I error**).
- The decision is to **not reject H_o** when, in fact, H_o is **false** (incorrect decision known as a **Type II error**).
- The decision is to **reject H_o** when, in fact, H_o is **false** (**correct decision** whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

α = probability of a Type I error = **P(Type I error)** = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = **P(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false.

α and β should be as small as possible because they are probabilities of errors. They are rarely 0.

The Power of the Test is $1 - \beta$. Ideally, we want a high power that is as close to 1 as possible. Increasing the sample size can increase the Power of the Test.

The following are examples of Type I and Type II errors.

Example:

Suppose the null hypothesis, H_0 , is: Frank's rock climbing equipment is safe.

Type I error: Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe. **Type II error:** Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.

α = **probability** that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe. β = **probability** that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.

Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)

Example:

Suppose the null hypothesis, H_0 , is: The victim of an automobile accident is alive when he arrives at the emergency room of a hospital.

Type I error: The emergency crew thinks that the victim is dead when, in fact, the victim is alive. **Type II error:** The emergency crew does not know if the victim is alive when, in fact, the victim is dead.

α = **probability** that the emergency crew thinks the victim is dead when, in fact, he is really alive = $P(\text{Type I error})$. β = **probability** that the

emergency crew does not know if the victim is alive when, in fact, the victim is dead = $P(\text{Type II error})$.

The error with the greater consequence is the Type I error. (If the emergency crew thinks the victim is dead, they will not treat him.)

Glossary

Type 1 Error

The decision is to reject the Null hypothesis when, in fact, the Null hypothesis is true.

Type 2 Error

The decision is to not reject the Null hypothesis when, in fact, the Null hypothesis is false.

Distribution Needed for Hypothesis Testing

Earlier in the course, we discussed sampling distributions. **Particular distributions are associated with hypothesis testing.** Perform tests of a population mean using a [normal distribution](#) or a [student's-t distribution](#). (Remember, use a student's-t distribution when the population [standard deviation](#) is unknown and the distribution of the sample mean is approximately normal.) In this chapter we perform tests of a population proportion using a normal distribution (usually n is large or the sample size is large).

If you are testing a **single population mean**, the distribution for the test is for **means**:

$$X \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right) \quad \text{or} \quad t_{df}$$

The population parameter is μ . The estimated value (point estimate) for μ is \bar{x} , the sample mean.

If you are testing a **single population proportion**, the distribution for the test is for proportions or percentages:

$$P' \sim N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$$

The population parameter is p . The estimated value (point estimate) for p is p' . $p' = \frac{x}{n}$ where x is the number of successes and n is the sample size.

Glossary

Normal Distribution

A continuous random variable (RV) with pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \text{ where } \mu \text{ is the mean of the distribution and}$$

σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Student's-t Distribution

Investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data.

Assumption

When you perform a **hypothesis test of a single population mean μ** using a **Student's-t distribution** (often called a t-test), there are fundamental assumptions that need to be met in order for the test to work properly. Your data should be a **simple random sample** that comes from a population that is approximately **normally distributed**. You use the sample **standard deviation** to approximate the population standard deviation. (Note that if the sample size is sufficiently large, a t-test will work even if the population is not approximately normally distributed).

When you perform a **hypothesis test of a single population mean μ** using a normal distribution (often called a z-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large. You know the value of the population standard deviation.

When you perform a **hypothesis test of a single population proportion p** , you take a simple random sample from the population. You must meet the conditions for a **binomial distribution** which are there are a certain number n of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success p . The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, the quantities np and nq must both be greater than five ($np > 5$ and $nq > 5$). Then the binomial distribution of sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{p \cdot q}{n}}$. Remember that $q = 1 - p$.

Glossary

Binomial Distribution

A discrete random variable (RV) which arises from Bernoulli trials. There are a fixed number, n , of independent trials. “Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is

defined as the number of successes in n trials. The notation is: $X \sim B(n, p)$. The mean is $\mu = np$ and the standard deviation is $\sigma = \sqrt{npq}$. The probability of exactly x successes in n trials is

$$P(X = x) = \binom{n}{x} p^x q^{n-x}.$$

Normal Distribution

A continuous random variable (RV) with pdf

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

where μ is the mean of the distribution and

σ is the standard deviation. Notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Student-t Distribution

Investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as n gets larger.
- There is a "family" of t distributions: every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data.

Rare Events

Suppose you make an assumption about a property of the population (this assumption is the [null hypothesis](#)). Then you gather sample data randomly. If the sample has properties that would be very **unlikely** to occur if the assumption is true, then you would conclude that your assumption about the population is probably incorrect. (Remember that your assumption is just an **assumption** - it is not a fact and it may or may not be true. But your sample data are real and the data are showing you a fact that seems to contradict your assumption.)

For example, Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a \$100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a \$100 bill. The probability of this happening is $\frac{1}{200} = 0.005$. Because this is so unlikely, Ali is hoping that what the two of them were told is wrong and there are more \$100 bills in the basket. A "rare event" has occurred (Didi getting the \$100 bill) so Ali doubts the assumption about only one \$100 bill being in the basket.

Glossary

Hypothesis

A statement about the value of a population parameter. In case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternate hypothesis (notation H_a).

Using the Sample to Support One of the Hypotheses

Use the sample data to calculate the actual probability of getting the test result, called the **p-value**. The p-value is the **probability that, if the null hypothesis is true, the results from another randomly selected sample will be as extreme or more extreme as the results obtained from the given sample.**

A large p-value calculated from the data indicates that we should fail to reject the **null hypothesis**. The smaller the p-value, the more unlikely the outcome, and the stronger the evidence is against the null hypothesis. We would reject the null hypothesis if the evidence is strongly against it.

Draw a graph that shows the p-value. The hypothesis test is easier to perform if you use a graph because you see the problem more clearly.

Example:

(to illustrate the p-value)

Suppose a baker claims that his bread height is more than 15 cm, on the average. Several of his customers do not believe him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The mean height of the sample loaves is 17 cm. The baker knows from baking hundreds of loaves of bread that the **standard deviation** for the height is 0.5 cm. and the distribution of heights is normal.

The null hypothesis could be $H_o: \mu \leq 15$ The alternate hypothesis is $H_a: \mu > 15$

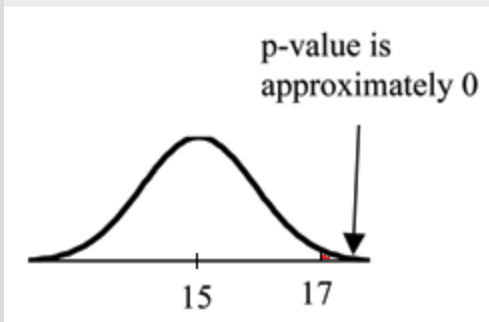
The words "**is more than**" translates as a ">" so " $\mu > 15$ " goes into the alternate hypothesis. The null hypothesis must contradict the alternate hypothesis.

Since σ **is known** ($\sigma = 0.5$ cm.), the distribution for the population is known to be normal with mean $\mu = 15$ and standard deviation $\frac{\sigma}{\sqrt{n}}$
 $= \frac{0.5}{\sqrt{10}} = 0.16.$

Suppose the null hypothesis is true (the mean height of the loaves is no more than 15 cm). Then is the mean height (17 cm) calculated from the

sample unexpectedly large? The hypothesis test works by asking the question how **unlikely** the sample mean would be if the null hypothesis were true. The graph shows how far out the sample mean is on the normal curve. The p-value is the probability that, if we were to take other samples, any other sample mean would fall at least as far out as 17 cm.

The p-value, then, is the probability that a sample mean is the same or greater than 17 cm. when the population mean is, in fact, 15 cm. We can calculate this probability using the normal distribution for means from Chapter 7.



$p\text{-value} = P(x > 17)$ which is approximately 0.

A p-value of approximately 0 tells us that it is highly unlikely that a loaf of bread rises no more than 15 cm, on the average. That is, almost 0% of all loaves of bread would be at least as high as 17 cm. **purely by CHANCE** had the population mean height really been 15 cm. Because the outcome of 17 cm. is so **unlikely (meaning it is happening NOT by chance alone)**, we conclude that the evidence is strongly against the null hypothesis (the mean height is at most 15 cm.). There is sufficient evidence that the true mean height for the population of the baker's loaves of bread is greater than 15 cm.

Glossary

Hypothesis

A statement about the value of a population parameter. In case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternate hypothesis (notation H_a).

p-value

The probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Decision and Conclusion

A systematic way to make a decision of whether to reject or not reject the **null hypothesis** is to compare the **p-value** and a **preset or preconceived α** (also called a "**significance level**"). A preset α is the probability of a **Type I error** (rejecting the null hypothesis when the null hypothesis is true). It may or may not be given to you at the beginning of the problem.

When you make a **decision** to reject or not reject H_o , do as follows:

- If $\alpha > \text{p-value}$, reject H_o . The results of the sample data are significant. There is sufficient evidence to conclude that H_o is an incorrect belief and that the **alternative hypothesis**, H_a , may be correct.
- If $\alpha \leq \text{p-value}$, do not reject H_o . The results of the sample data are not significant. There is not sufficient evidence to conclude that the alternative hypothesis, H_a , may be correct.
- When you "do not reject H_o ", it does not mean that you should believe that H_o is true. It simply means that the sample data have **failed** to provide sufficient evidence to cast serious doubt about the truthfulness of H_o .

Conclusion: After you make your decision, write a thoughtful **conclusion** about the hypotheses in terms of the given problem.

Glossary

Hypothesis

A statement about the value of a population parameter. In case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternate hypothesis (notation H_a).

Level of Significance of the Test

Probability of a Type I error (reject the null hypothesis when it is true).
Notation: α . In hypothesis testing, the Level of Significance is called the preconceived α or the preset α .

p-value

The probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

Type 1 Error

The decision is to reject the Null hypothesis when, in fact, the Null hypothesis is true.

Additional Information

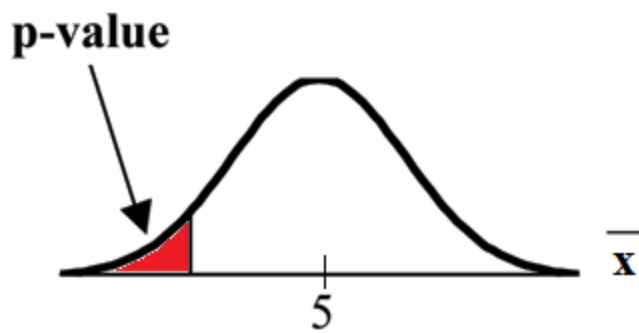
- In a **hypothesis test** problem, you may see words such as "the level of significance is 1%." The "1%" is the preconceived or preset α .
- The statistician setting up the hypothesis test selects the value of α to use **before** collecting the sample data.
- **If no level of significance is given, the accepted standard is to use $\alpha = 0.05$.**
- When you calculate the **p-value** and draw the picture, the p-value is the area in the left tail, the right tail, or split evenly between the two tails. For this reason, we call the hypothesis test left, right, or two tailed.
- The **alternate hypothesis**, H_a , tells you if the test is left, right, or two-tailed. It is the **key** to conducting the appropriate test.
- H_a **never** has a symbol that contains an equal sign.
- **Thinking about the meaning of the p-value:** A data analyst (and anyone else) should have more confidence that he made the correct decision to reject the null hypothesis with a smaller p-value (for example, 0.001 as opposed to 0.04) even if using the 0.05 level for alpha. Similarly, for a large p-value like 0.4, as opposed to a p-value of 0.056 (alpha = 0.05 is less than either number), a data analyst should have more confidence that she made the correct decision in failing to reject the null hypothesis. This makes the data analyst use judgment rather than mindlessly applying rules.

The following examples illustrate a left, right, and two-tailed test.

Example:

$$H_o: \mu = 5 \quad H_a: \mu < 5$$

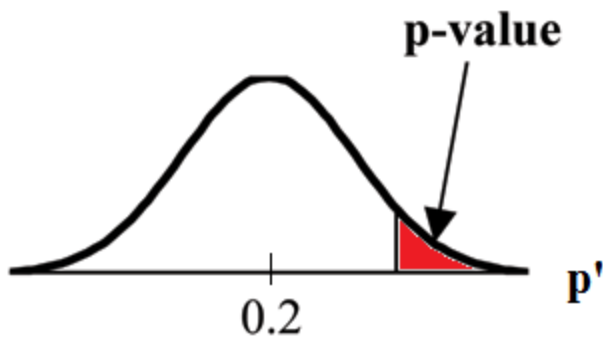
Test of a single population mean. H_a tells you the test is left-tailed. The picture of the p-value is as follows:



Example:

$$H_o: p \leq 0.2 \quad H_a: p > 0.2$$

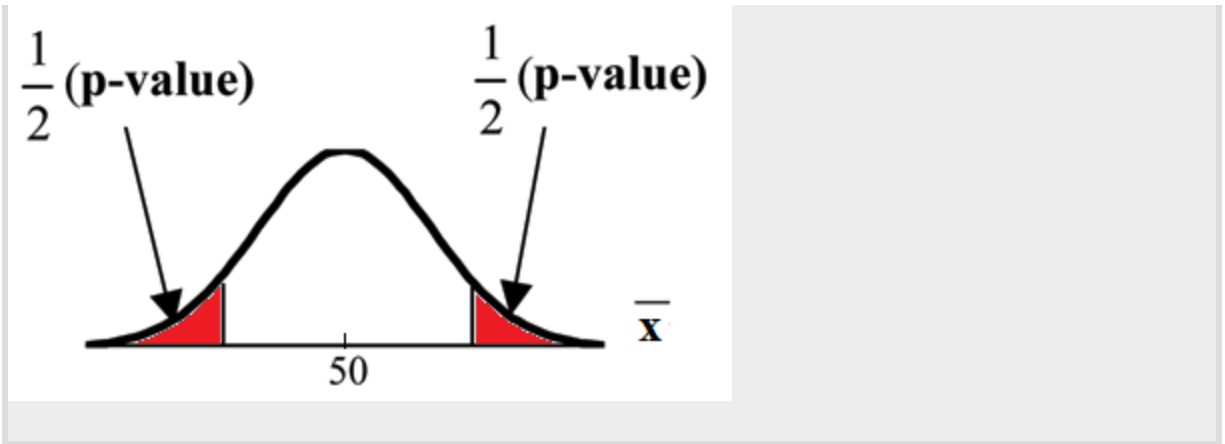
This is a test of a single population proportion. H_a tells you the test is **right-tailed**. The picture of the p-value is as follows:



Example:

$$H_o: \mu = 50 \quad H_a: \mu \neq 50$$

This is a test of a single population mean. H_a tells you the test is **two-tailed**. The picture of the p-value is as follows.



Glossary

Hypothesis Testing

Based on sample evidence, a procedure to determine whether the hypothesis stated is a reasonable statement and cannot be rejected, or is unreasonable and should be rejected.

p-value

The probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

Summary of the Hypothesis Test

The **hypothesis test** itself has an established process. This can be summarized as follows:

1. Determine H_o and H_a . Remember, they are contradictory.
2. Determine the random variable.
3. Determine the distribution for the test.
4. Draw a graph, calculate the test statistic, and use the test statistic to calculate the **p-value**. (A z-score and a t-score are examples of test statistics.)
5. Compare the preconceived α with the p-value, make a decision (reject or do not reject H_o), and write a clear conclusion using English sentences.

Notice that in performing the hypothesis test, you use α and not β . β is needed to help determine the sample size of the data that is used in calculating the p-value. Remember that the quantity $1 - \beta$ is called the **Power of the Test**. A high power is desirable. If the power is too low, statisticians typically increase the sample size while keeping α the same. If the power is low, the null hypothesis might not be rejected when it should be.

Glossary

Hypothesis Testing

Based on sample evidence, a procedure to determine whether the hypothesis stated is a reasonable statement and cannot be rejected, or is unreasonable and should be rejected.

p-value

The probability that an event will happen purely by chance assuming the null hypothesis is true. The smaller the p-value, the stronger the evidence is against the null hypothesis.

Examples

This module provides examples of Hypothesis Testing of a Single Mean and a Single Proportion as a part of the Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Example:

Exercise:

Problem:

Jeffrey, as an eight-year old, **established a mean time of 16.43 seconds** for swimming the 25-yard freestyle, with a **standard deviation of 0.8 seconds**. His dad, Frank, thought that Jeffrey could swim the 25-yard freestyle faster by using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey for **15 25-yard freestyle swims**. For the 15 swims, **Jeffrey's mean time was 16 seconds**. **Frank thought that the goggles helped Jeffrey to swim faster than the 16.43 seconds**. Conduct a hypothesis test using a preset $\alpha = 0.05$. Assume that the swim times for the 25-yard freestyle are normal.

Solution:

Set up the Hypothesis Test:

Since the problem is about a mean, this is a **test of a single population mean**.

$$H_o: \mu = 16.43 \quad H_a: \mu < 16.43$$

For Jeffrey to swim faster, his time will be less than 16.43 seconds. The "<" tells you this is left-tailed.

Determine the distribution needed:

Random variable: X = the mean time to swim the 25-yard freestyle.

Distribution for the test: X is normal (population standard deviation is known: $\sigma = 0.8$)

$$X \sim N\left(\mu, \frac{\sigma_X}{\sqrt{n}}\right) \quad \text{Therefore, } X \sim N\left(16.43, \frac{0.8}{\sqrt{15}}\right)$$

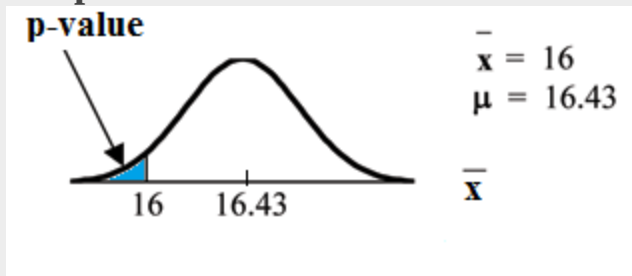
$\mu = 16.43$ comes from H_0 and not the data. $\sigma = 0.8$, and $n = 15$.

Calculate the p-value using the normal distribution for a mean:

p-value = $P(\bar{x} < 16) = 0.0187$ where the sample mean in the problem is given as 16.

p-value = 0.0187 (This is called the **actual level of significance**.)
The p-value is the area to the left of the sample mean is given as 16.

Graph:



$\mu = 16.43$ comes from H_0 . Our assumption is $\mu = 16.43$.

Interpretation of the p-value: If H_0 is true, there is a 0.0187 probability (1.87%) that Jeffrey's mean time to swim the 25-yard freestyle is 16 seconds or less. Because a 1.87% chance is small, the mean time of 16 seconds or less is unlikely to have happened randomly. It is a rare event.

Compare α and the p-value:

$$\alpha = 0.05 \quad \text{p-value} = 0.0187 \quad \alpha > \text{p-value}$$

Make a decision: Since $\alpha > \text{p-value}$, reject H_0 .

This means that you reject $\mu = 16.43$. In other words, you do not think Jeffrey swims the 25-yard freestyle in 16.43 seconds but faster with the new goggles.

Conclusion: At the 5% significance level, we conclude that Jeffrey swims faster using the new goggles. The sample data show there is sufficient evidence that Jeffrey's mean time to swim the 25-yard freestyle is less than 16.43 seconds.

The p-value can easily be calculated using the TI-83+ and the TI-84 calculators:

Press **STAT** and arrow over to **TESTS**. Press **1:Z-Test**. Arrow over to **Stats** and press **ENTER**. Arrow down and enter 16.43 for μ_0 (null hypothesis), .8 for σ , 16 for the sample mean, and 15 for n . Arrow down to μ : (alternate hypothesis) and arrow over to $<\mu_0$. Press **ENTER**. Arrow down to **Calculate** and press **ENTER**. The calculator not only calculates the p-value ($p = 0.0187$) but it also calculates the test statistic (z-score) for the sample mean. $\mu < 16.43$ is the alternate hypothesis. Do this set of instructions again except arrow to **Draw** (instead of **Calculate**). Press **ENTER**. A shaded graph appears with $z = -2.08$ (test statistic) and $p = 0.0187$ (p-value). Make sure when you use **Draw** that no other equations are highlighted in $Y =$ and the plots are turned off.

When the calculator does a Z-Test, the **Z-Test** function finds the p-value by doing a normal probability calculation using the **Central Limit Theorem**:

$$P(x < 16) = \text{2nd DISTR normcdf} \left(-10 \wedge 99, 16, 16.43, 0.8/\sqrt{15} \right).$$

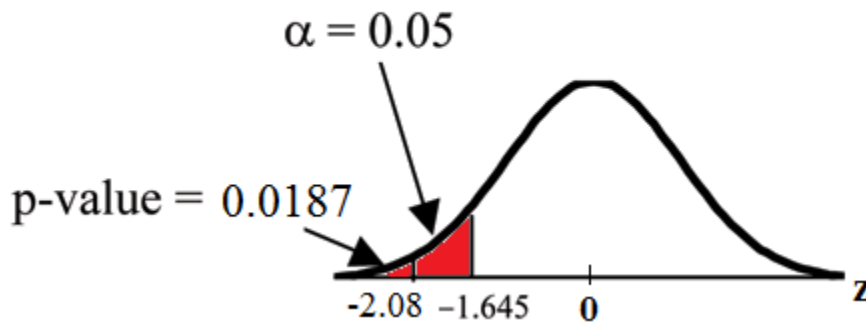
The Type I and Type II errors for this problem are as follows:

The Type I error is to conclude that Jeffrey swims the 25-yard freestyle, on average, in less than 16.43 seconds when, in fact, he

actually swims the 25-yard freestyle, on average, in 16.43 seconds.
(Reject the null hypothesis when the null hypothesis is true.)

The Type II error is that there is not evidence to conclude that Jeffrey swims the 25-yard free-style, on average, in less than 16.43 seconds when, in fact, he actually does swim the 25-yard free-style, on average, in less than 16.43 seconds. (Do not reject the null hypothesis when the null hypothesis is false.)

Historical Note: The traditional way to compare the two probabilities, α and the p-value, is to compare the critical value (z-score from α) to the test statistic (z-score from data). The calculated test statistic for the p-value is -2.08 . (From the Central Limit Theorem, the test statistic formula is $z = \frac{x - \mu_X}{\left(\frac{\sigma_X}{\sqrt{n}}\right)}$. For this problem, $x = 16$, $\mu_X = 16.43$ from the null hypothesis, $\sigma_X = 0.8$, and $n = 15$.) You can find the critical value for $\alpha = 0.05$ in the normal table (see **15.Tables** in the Table of Contents). The z-score for an area to the left equal to 0.05 is midway between -1.65 and -1.64 (0.05 is midway between 0.0505 and 0.0495). The z-score is -1.645 . Since $-1.645 > -2.08$ (which demonstrates that $\alpha > \text{p-value}$), reject H_0 . Traditionally, the decision to reject or not reject was done in this way. Today, comparing the two probabilities α and the p-value is very common. For this problem, the p-value, 0.0187 is considerably smaller than α , 0.05 . You can be confident about your decision to reject. The graph shows α , the p-value, and the test statistics and the critical value.



Example:**Exercise:****Problem:**

A college football coach thought that his players could bench press a **mean weight of 275 pounds**. It is known that the **standard deviation is 55 pounds**. Three of his players thought that the mean weight was **more than** that amount. They asked **30** of their teammates for their estimated maximum lift on the bench press exercise. The data ranged from 205 pounds to 385 pounds. The actual different weights were (frequencies are in parentheses) 205(3) 215(3) 225(1) 241(2) 252(2) 265(2) 275(2) 313(2) 316(5) 338(2) 341(1) 345(2) 368(2) 385(1). (Source: data from Reuben Davis, Kraig Evans, and Scott Gunderson.)

Conduct a hypothesis test using a 2.5% level of significance to determine if the bench press mean is **more than 275 pounds**.

Solution:

Set up the Hypothesis Test:

Since the problem is about a mean weight, this is a **test of a single population mean**.

$H_0: \mu = 275$ $H_a: \mu > 275$ This is a right-tailed test.

Calculating the distribution needed:

Random variable: X = the mean weight, in pounds, lifted by the football players.

Distribution for the test: It is normal because σ is known.

$$X \sim N \left(275, \frac{55}{\sqrt{30}} \right)$$

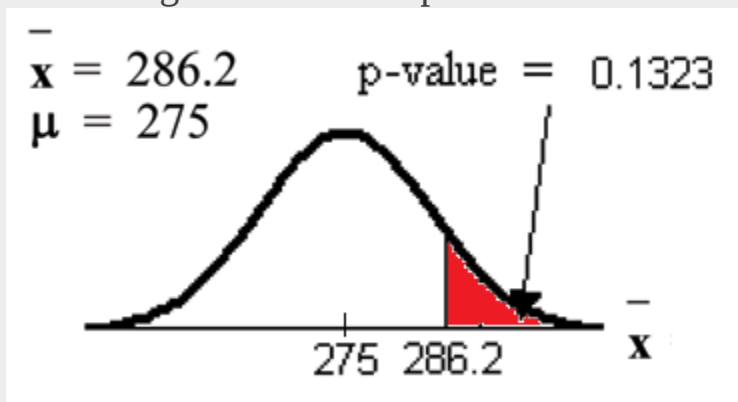
$x = 286.2$ pounds (from the data).

$\sigma = 55$ pounds (**Always use σ if you know it.**) We assume $\mu = 275$ pounds unless our data shows us otherwise.

Calculate the p-value using the normal distribution for a mean and using the sample mean as input (see the calculator instructions below for using the data as input):

$$\text{p-value} = P(x > 286.2) = 0.1323.$$

Interpretation of the p-value: If H_0 is true, then there is a 0.1331 probability (13.23%) that the football players can lift a mean weight of 286.2 pounds or more. Because a 13.23% chance is large enough, a mean weight lift of 286.2 pounds or more is not a rare event.



Compare α and the p-value:

$$\alpha = 0.025 \quad \text{p-value} = 0.1323$$

Make a decision: Since $\alpha < \text{p-value}$, do not reject H_0 .

Conclusion: At the 2.5% level of significance, from the sample data, there is not sufficient evidence to conclude that the true mean weight lifted is more than 275 pounds.

The p-value can easily be calculated using the TI-83+ and the TI-84 calculators:

Put the data and frequencies into lists. Press **STAT** and arrow over to **TESTS**. Press **1:Z-Test**. Arrow over to **Data** and press **ENTER**.

Arrow down and enter 275 for μ_0 , 55 for σ , the name of the list where you put the data, and the name of the list where you put the frequencies. Arrow down to $\mu :$ and arrow over to $> \mu_0$. Press **ENTER**. Arrow down to **Calculate** and press **ENTER**. The calculator not only calculates the p-value ($p = 0.1331$, a little different from the above calculation - in it we used the sample mean rounded to one decimal place instead of the data) but it also calculates the test statistic (z-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 275$ is the alternate hypothesis. Do this set of instructions again except arrow to **Draw** (instead of **Calculate**). Press **ENTER**. A shaded graph appears with $z = 1.112$ (test statistic) and $p = 0.1331$ (p-value). Make sure when you use **Draw** that no other equations are highlighted in $Y =$ and the plots are turned off.

Example:

Exercise:

Problem:

Statistics students believe that the mean score on the first statistics test is 65. A statistics instructor thinks the mean score is higher than 65. He samples ten statistics students and obtains the scores 65 65 70 67 66 63 63 68 72 71. He performs a hypothesis test using a 5% level of significance. The data are from a normal distribution.

Solution:

Set up the Hypothesis Test:

A 5% level of significance means that $\alpha = 0.05$. This is a test of a **single population mean**.

$$H_o: \mu = 65 \quad H_a: \mu > 65$$

Since the instructor thinks the average score is higher, use a ">". The ">" means the test is right-tailed.

Determine the distribution needed:

Random variable: \bar{X} = average score on the first statistics test.

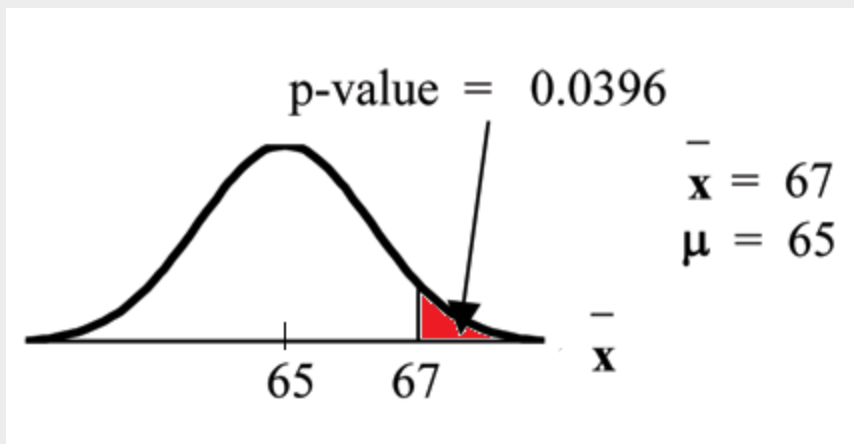
Distribution for the test: If you read the problem carefully, you will notice that there is **no population standard deviation given**. You are only given $n = 10$ sample data values. Notice also that the data come from a normal distribution. This means that the distribution for the test is a student's-t.

Use t_{df} . Therefore, the distribution for the test is t_9 where $n = 10$ and $df = 10 - 1 = 9$.

Calculate the p-value using the Student's-t distribution:

p-value = $P(x > 67) = 0.0396$ where the sample mean and sample standard deviation are calculated as 67 and 3.1972 from the data.

Interpretation of the p-value: If the null hypothesis is true, then there is a 0.0396 probability (3.96%) that the sample mean is 67 or more.



Compare α and the p-value:

Since $\alpha = .05$ and p-value = 0.0396. Therefore, $\alpha > \text{p-value}$.

Make a decision: Since $\alpha > \text{p-value}$, reject H_0 .

This means you reject $\mu = 65$. In other words, you believe the average test score is more than 65.

Conclusion: At a 5% level of significance, the sample data show sufficient evidence that the mean (average) test score is more than 65, just as the math instructor thinks.

The p-value can easily be calculated using the TI-83+ and the TI-84 calculators:

Put the data into a list. Press **STAT** and arrow over to **TESTS**. Press **2:T-Test**. Arrow over to **Data** and press **ENTER**. Arrow down and enter 65 for μ_0 , the name of the list where you put the data, and 1 for **Freq:**. Arrow down to $\mu :$ and arrow over to $> \mu_0$. Press **ENTER**. Arrow down to **Calculate** and press **ENTER**. The calculator not only calculates the p-value ($p = 0.0396$) but it also calculates the test statistic (t-score) for the sample mean, the sample mean, and the sample standard deviation. $\mu > 65$ is the alternate hypothesis. Do this set of instructions again except arrow to **Draw** (instead of **Calculate**). Press **ENTER**. A shaded graph appears with $t = 1.9781$ (test statistic) and $p = 0.0396$ (p-value). Make sure when you use **Draw** that no other equations are highlighted in $Y =$ and the plots are turned off.

Example:

Exercise:

Problem:

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is **the same or different from 50%**. Joon samples **100 first-time brides** and **53** reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

Solution:

Set up the Hypothesis Test:

The 1% level of significance means that $\alpha = 0.01$. This is a **test of a single population proportion**.

$$H_o: p = 0.50 \quad H_a: p \neq 0.50$$

The words "**is the same or different from**" tell you this is a two-tailed test.

Calculate the distribution needed:

Random variable: $P\%$ = the percent of first-time brides who are younger than their grooms.

Distribution for the test: The problem contains no mention of a mean. The information is given in terms of percentages. Use the distribution for $P\%$, the estimated proportion.

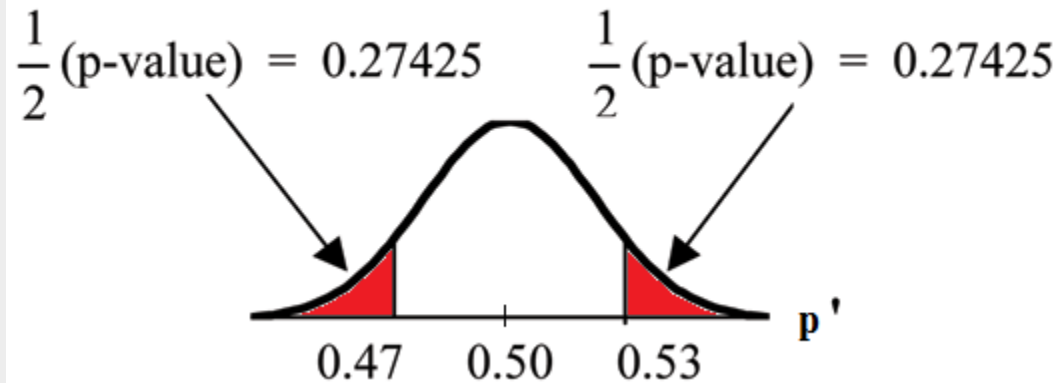
$$P\% \sim N \left(p, \sqrt{\frac{p \cdot q}{n}} \right) \quad \text{Therefore, } P\% \sim N \left(0.5, \sqrt{\frac{0.5 \cdot 0.5}{100}} \right) \text{ where } p = 0.50, q = 1 - p = 0.50, \text{ and } n = 100.$$

Calculate the p-value using the normal distribution for proportions:

$$\text{p-value} = P(p' < 0.47 \text{ or } p' > 0.53) = 0.5485$$

where $x = 53$, $p' = \frac{x}{n} = \frac{53}{100} = 0.53$.

Interpretation of the p-value: If the null hypothesis is true, there is 0.5485 probability (54.85%) that the sample (estimated) proportion p' is 0.53 or more OR 0.47 or less (see the graph below).



$\mu = p = 0.50$ comes from H_0 , the null hypothesis.

$p' = 0.53$. Since the curve is symmetrical and the test is two-tailed, the p' for the left tail is equal to $0.50 - 0.03 = 0.47$ where $\mu = p = 0.50$. (0.03 is the difference between 0.53 and 0.50.)

Compare α and the p-value:

Since $\alpha = 0.01$ and p-value = 0.5485. Therefore, $\alpha < \text{p-value}$.

Make a decision: Since $\alpha < \text{p-value}$, you cannot reject H_0 .

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of first-time brides that are younger than their grooms is different from 50%.

The p-value can easily be calculated using the TI-83+ and the TI-84 calculators:

Press **STAT** and arrow over to **TESTS**. Press **5:1-PropZTest**. Enter .5 for p_0 , 53 for x and 100 for n . Arrow down to **Prop** and arrow to **not equals** p_0 . Press **ENTER**. Arrow down to

Calculate and press **ENTER**. The calculator calculates the p-value ($p = 0.5485$) and the test statistic (z-score). **Prop not equals .5** is the alternate hypothesis. Do this set of instructions again except arrow to **Draw** (instead of **Calculate**). Press **ENTER**. A shaded graph appears with $z = 0.6$ (test statistic) and $p = 0.5485$ (p-value). Make sure when you use **Draw** that no other equations are highlighted in $Y =$ and the plots are turned off.

The Type I and Type II errors are as follows:

The Type I error is to conclude that the proportion of first-time brides that are younger than their grooms is different from 50% when, in fact, the proportion is actually 50%. (Reject the null hypothesis when the null hypothesis is true).

The Type II error is there is not enough evidence to conclude that the proportion of first time brides that are younger than their grooms differs from 50% when, in fact, the proportion does differ from 50%. (Do not reject the null hypothesis when the null hypothesis is false.)

Example:

Exercise:

Problem:

Suppose a consumer group suspects that the proportion of households that have three cell phones is 30%. A cell phone company has reason to believe that the proportion is 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three cell phones.

Solution:

Set up the Hypothesis Test:

$$H_o: p = 0.30 \quad H_a: p \neq 0.30$$

Determine the distribution needed:

The **random variable** is P' = proportion of households that have three cell phones.

The **distribution** for the hypothesis test is $P' \sim N$

$$\left(0.30, \sqrt{\frac{(0.30) \cdot (0.70)}{150}} \right)$$

Exercise:

Problem:

The value that helps determine the p-value is p' . Calculate p' .

Solution:

$p' = \frac{x}{n}$ where x is the number of successes and n is the total number in the sample.

$$x = 43, n = 150$$

$$p' = \frac{43}{150}$$

Exercise:

Problem: What is a **success** for this problem?

Solution:

A success is having three cell phones in a household.

Exercise:

Problem: What is the level of significance?

Solution:

The level of significance is the preset α . Since α is not given, assume that $\alpha = 0.05$.

Draw the graph for this problem. Draw the horizontal axis. Label and shade appropriately.

Exercise:

Problem: Calculate the p-value.

Solution:

p-value = 0.7216

Exercise:

Problem:

Make a decision. _____ (Reject/Do not reject) H_0
because _____.

Solution:

Assuming that $\alpha = 0.05$, $\alpha < \text{p-value}$. The Decision is do not reject H_0 because there is not sufficient evidence to conclude that the proportion of households that have three cell phones is not 30%.

The next example is a poem written by a statistics student named Nicole Hart. The solution to the problem follows the poem. Notice that the hypothesis test is for a single population proportion. This means that the null and alternate hypotheses use the parameter p . The distribution for the test is normal. The estimated proportion p' is the proportion of fleas killed to the total fleas found on Fido. This is sample information. The problem gives a preconceived $\alpha = 0.01$, for comparison, and a 95% confidence interval computation. The poem is clever and humorous, so please enjoy it!

Note: Hypothesis testing problems consist of multiple steps. To help you do the problems, solution sheets are provided for your use. Look in the Table of Contents Appendix for the topic "Solution Sheets." If you like, use copies of the appropriate solution sheet for homework problems.

Example:

Exercise:

Problem:

My dog has so many fleas, They do not come off with ease. As for shampoo, I have tried many types Even one called Bubble Hype, Which only killed 25% of the fleas, Unfortunately I was not pleased. I've used all kinds of soap, Until I had give up hope Until one day I saw An ad that put me in awe. A shampoo used for dogs Called GOOD ENOUGH to Clean a Hog Guaranteed to kill more fleas. I gave Fido a bath And after doing the math His number of fleas Started dropping by 3's! Before his shampoo I counted 42. At the end of his bath, I redid the math And the new shampoo had killed 17 fleas. So now I was pleased. Now it is time for you to have some fun With the level of significance being .01, You must help me figure out Use the new shampoo or go without?

Solution:

Set up the Hypothesis Test:

$$H_o: p = 0.25 \quad H_a: p > 0.25$$

Determine the distribution needed:

In words, CLEARLY state what your random variable X or P' represents.

P' = The proportion of fleas that are killed by the new shampoo

State the distribution to use for the test.

Normal: $N\left(0.25, \sqrt{\frac{(0.25)(1-0.25)}{42}}\right)$

Test Statistic: $z = 2.3163$

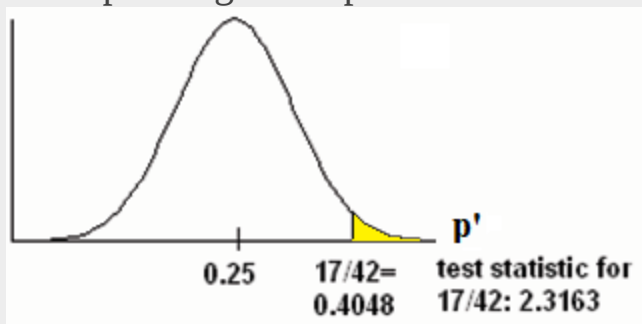
Calculate the p-value using the normal distribution for proportions:

p-value = 0.0103

In 1 – 2 complete sentences, explain what the p-value means for this problem.

If the null hypothesis is true (the proportion is 0.25), then there is a 0.0103 probability that the sample (estimated) proportion is 0.4048 ($\frac{17}{42}$) or more.

Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the p-value.



Compare α and the p-value:

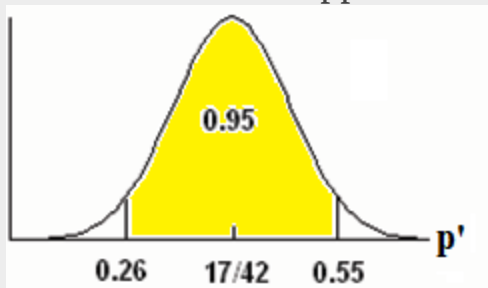
Indicate the correct decision (“reject” or “do not reject” the null hypothesis), the reason for it, and write an appropriate conclusion,

using COMPLETE SENTENCES.

alpha	decision	reason for decision
0.01	Do not reject H_o	$\alpha < p\text{-value}$

Conclusion: At the 1% level of significance, the sample data do not show sufficient evidence that the percentage of fleas that are killed by the new shampoo is more than 25%.

Construct a 95% Confidence Interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the Confidence Interval.



Confidence Interval: (0.26, 0.55) We are 95% confident that the true population proportion p of fleas that are killed by the new shampoo is between 26% and 55%.

Note: This test result is not very definitive since the p-value is very close to alpha. In reality, one would probably do more tests by giving the dog another bath after the fleas have had a chance to return.

Glossary

Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation σ . We are sampling with size n and we are interested in two new RVs - the sample mean, \bar{X} , and the sample sum, ΣX . If the size n of the sample is sufficiently large, then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and $\Sigma X \sim N(n\mu, \sqrt{n}\sigma)$. If the size n of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distribution regardless of the shape of the population. The mean of the sample means will equal the population mean and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Summary of Formulas

H_o and H_a are contradictory.

If H_o has:	equal	greater than or equal to	less than or equal to
then H_a has:	not equal than or greater or less than	less than	greater than

If α = p-value, then do not reject H_o .

If α = p-value, then reject H_o .

α is preconceived. Its value is set before the hypothesis test starts. The p-value is calculated from the data.

α = probability of a Type I error = P(Type I error) = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = P(Type II error) = probability of not rejecting the null hypothesis when the null hypothesis is false.

If there is no given preconceived α , then use α .

Types of Hypothesis Tests

- Single population mean, **known** population variance (or standard deviation): **Normal test**.
- Single population mean, **unknown** population variance (or standard deviation): **Student's-t test**.
- Single population proportion: **Normal test**.

Effect Size

A brief discussion of Effect Size for an intro statistics class.

Effect Size

When we say that a finding is statistically significant we must remember that we mean only that it probably did not occur by chance. However, saying that a finding is statistically significant does not tell us if the observed result is meaningful or important as in the more general usage of the word significant.

Researchers use the term **effect size** to measure the strength of the relationship between two variables.

There are many different effect sizes in use. Different statistical analyses also use different measure of effect size.

The most popular measure of effect size for t-tests is Cohen's *d*.

For t-test with independent samples Cohen's *d* = mean of experimental group – mean of control group divided by the standard deviation of control group.

$$d = \frac{\bar{Y}_a - \bar{Y}_b}{S_b}$$

For repeated measures t-test researchers often use Coehn's *d* with the mean of the pre-test minus the mean of the post-test divided by the mean of the pre-test.

$$d = \frac{\bar{Y}_{pre} - \bar{Y}_{post}}{S_{pre}}$$

For single sample t-tests Choen's d is equal to the mean of the sample minus the hypothesized population mean divided by the standard deviation of the sample.

$$d = \frac{M - \mu}{S}$$

Values of estimates of effect size

Effect Size	Description of Effect
Cohen's d = less than .2	Small
Cohen's d is between .2 and .8	Medium
Cohen's d is greater than .8	Large

For correlations we use the Coefficient of Determination or R^2 as a measure of effect size.

Values of estimates of effect size, r^2

Effect Size	Description of Effect
r^2 = less than .01	Small
r^2 = less than .09	Medium
r^2 = less than .25	Large

Cohen, J. (1988). *Statistical Power Analysis for the Behavioal Sciences*. Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, 112, 155-159.

Grissom, R. J. & Kim, J. J. (2005). *Effect Sizes for Research*. Mahway, NJ: Erlbaum.

Hypothesis Testing: Two Population Means and Two Population Proportions

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Classify hypothesis tests by type.
- Conduct and interpret hypothesis tests for two population means, population standard deviations known.
- Conduct and interpret hypothesis tests for two population means, population standard deviations unknown.
- Conduct and interpret hypothesis tests for two population proportions.
- Conduct and interpret hypothesis tests for matched or paired samples.

Introduction

Studies often compare two groups. For example, researchers are interested in the effect aspirin has in preventing heart attacks. Over the last few years, newspapers and magazines have reported about various aspirin studies involving two groups. Typically, one group is given aspirin and the other group is given a placebo. Then, the heart attack rate is studied over several years.

There are other situations that deal with the comparison of two groups. For example, studies compare various diet and exercise programs. Politicians compare the proportion of individuals from different income brackets who might vote for them. Students are interested in whether SAT or GRE preparatory courses really help raise their scores.

In the previous chapter, you learned to conduct hypothesis tests on single means and single proportions. You will expand upon that in this chapter. You will compare two means or two proportions to each other. The general procedure is still the same, just expanded.

To compare two means or two proportions, you work with two groups. The groups are classified either as **independent** or **matched pairs**.

Independent groups mean that the two samples taken are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. **Matched pairs** consist of two samples that are dependent. The parameter tested using matched pairs is the population mean. The parameters tested using independent groups are either population means or population proportions.

Note: This chapter relies on either a calculator or a computer to calculate the degrees of freedom, the test statistics, and p-values. TI-83+ and TI-84 instructions are included as well as the test statistic formulas. When using the TI-83+/TI-84 calculators, we do not need to separate two population means, independent groups, population variances unknown into large and small sample sizes. However, most statistical computer software has the ability to differentiate these tests.

This chapter deals with the following hypothesis tests:

Independent groups (samples are independent)

- Test of two population means.
- Test of two population proportions.

Matched or paired samples (samples are dependent)

- Becomes a test of one population mean.

Comparing Two Independent Population Means with Unknown Population Standard Deviations

This module provides an overview of Comparing Two Independent Population Means with Unknown Population Standard Deviations as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

1. The two independent samples are simple random samples from two distinct populations.
2. Both populations are normally distributed with the population means and standard deviations unknown unless the sample sizes are greater than 30. In that case, the populations need not be normally distributed.

Note: The test comparing two independent population means with unknown and possibly unequal population standard deviations is called the Aspin-Welch t-test. The degrees of freedom formula was developed by Aspin-Welch.

The comparison of two population means is very common. A difference between the two samples depends on both the means and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples. In order to account for the variation, we take the difference of the sample means, $\bar{x}_1 - \bar{x}_2$, and divide by the standard error (shown below) in order to standardize the difference. The result is a t-score test statistic (shown below).

Because we do not know the population standard deviations, we estimate them using the two sample standard deviations from our independent samples. For the hypothesis test, we calculate the estimated standard deviation, or **standard error**, of the difference in sample means, $\sqrt{s_1^2/n_1 + s_2^2/n_2}$.

Equation:

The standard error is:

$$\frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{1}{n_1} + \frac{1}{n_2}}$$

The test statistic (t-score) is calculated as follows:

Equation:

t-score

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where:

- s_1 and s_2 , the sample standard deviations, are estimates of σ_1 and σ_2 , respectively.
- σ_1 and σ_2 are the unknown population standard deviations.
- \bar{x}_1 and \bar{x}_2 are the sample means. μ_1 and μ_2 are the population means.

The **degrees of freedom (df)** is a somewhat complicated calculation. However, a computer or calculator calculates it easily. The dfs are not always a whole number. The test statistic calculated above is approximated by the student's-t distribution with dfs as follows:

Equation:

Degrees of freedom

$$\frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{1}{n_1} + \frac{1}{n_2}}$$

When both sample sizes n_1 and n_2 are five or larger, the student's-t approximation is very good. Notice that the sample variances s_1^2 and s_2^2 are not pooled. (If the question comes up, do not pool the variances.)

Note: It is not necessary to compute this by hand. A calculator or computer easily computes it.

Example:

Independent groups

The average amount of time boys and girls ages 7 through 11 spend playing sports each day is believed to be the same. An experiment is done, data is collected, resulting in the table below. Both populations have a normal distribution.

	Sample Size	Average Number of Hours Playing Sports Per Day	Sample Standard Deviation
Girls	9	2 hours	—
Boys	16	3.2 hours	1.00

Exercise:

Problem:

Is there a difference in the mean amount of time boys and girls ages 7 through 11 play sports each day? Test at the 5% level of significance.

Solution:

The population standard deviations are not known. Let μ_1 be the subscript for girls and μ_2 be the subscript for boys. Then, $\mu_1 - \mu_2$ is the

population mean for girls and is the population mean for boys.
This is a test of two **independent groups**, two population **means**.

Random variable: = difference in the sample mean amount
of time girls and boys play sports each day.

:

:

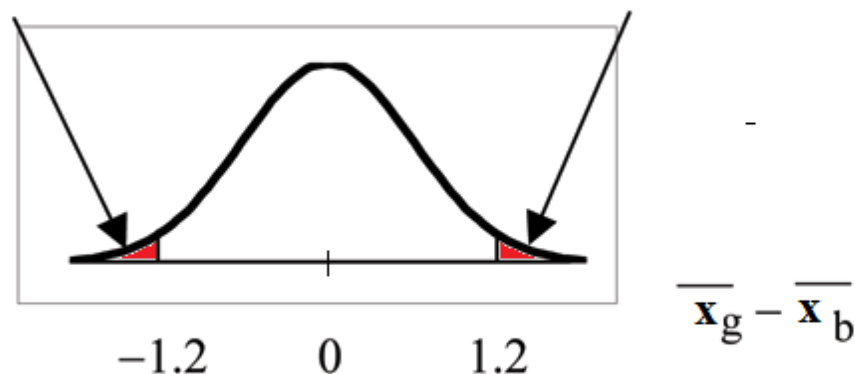
The words "**the same**" tell you has an "=". Since there are no
other words to indicate , then assume "**is different**." This is a two-
tailed test.

Distribution for the test: Use where is calculated using the
formula for independent groups, two population means. Using a
calculator, is approximately 18.8462. **Do not pool the variances.**

Calculate the p-value using a student's-t distribution: p-value =
0.0054

Graph:

$$\frac{1}{2} (\text{p-value}) = 0.0028 \qquad \frac{1}{2} (\text{p-value}) = 0.0028$$



From H_0 , $\mu_g - \mu_b = 0$

So,

Half the p-value is below -1.2 and half is above 1.2.

Make a decision: Since p-value, reject .

This means you reject . The means are different.

Conclusion: At the 5% level of significance, the sample data show there is sufficient evidence to conclude that the mean number of hours that girls and boys aged 7 through 11 play sports per day is different (mean number of hours boys aged 7 through 11 play sports per day is greater than the mean number of hours played by girls OR the mean number of hours girls aged 7 through 11 play sports per day is greater than the mean number of hours played by boys).

Note:TI-83+ and TI-84: Press

STAT

. Arrow over to

TESTS

and press

4:2-SampTTest

. Arrow over to Stats and press

ENTER

. Arrow down and enter

2

for the first sample mean,

for Sx_1 ,

9

for n_1 ,

3.2

for the second sample mean,

1

for Sx_2 , and

16

for n_2 . Arrow down to μ_1 : and arrow to

does not equal

μ_2 . Press

ENTER

. Arrow down to Pooled: and

No

. Press

ENTER

. Arrow down to

Calculate

and press

ENTER

. The p-value is $p = 0.0054$, the dfs are approximately 18.8462, and the test statistic is -3.14. Do the procedure again but instead of Calculate do Draw.

Example:

A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples 11 graduates. Their average is 4 math classes with a standard deviation of 1.5 math classes. College B samples 9 graduates. Their average is 3.5 math classes with a standard deviation of 1 math class. The community group believes that a student who graduates from college A **has taken more math classes**, on the average. Both populations have a normal distribution. Test at a 1% significance level. Answer the following questions.

Exercise:

Problem: Is this a test of two means or two proportions?

Solution:

two means

Exercise:

Problem:

Are the populations standard deviations known or unknown?

Solution:

unknown

Exercise:

Problem: Which distribution do you use to perform the test?

Solution:

student's-t

Exercise:

Problem: What is the random variable?

Solution:**Exercise:**

Problem: What are the null and alternate hypothesis?

Solution:

-
-

Exercise:

Problem: Is this test right, left, or two tailed?

Solution:

right

Exercise:

Problem: What is the p-value?

Solution:

0.1928

Exercise:

Problem: Do you reject or not reject the null hypothesis?

Solution:

Do not reject.

Conclusion:

At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a student who graduates from college A has taken more math classes, on the average, than a student who graduates from college B.

Glossary

Degrees of Freedom (df)

The number of objects in a sample that are free to vary.

Standard Deviation

A number that is equal to the square root of the variance and measures how far data values are from their mean. Notation: s for sample standard deviation and σ for population standard deviation.

Variable (Random Variable)

A characteristic of interest in a population being studied. Common notation for variables are upper case Latin letters X, Y, Z, \dots ; common notation for a specific value from the domain (set of all possible values of a variable) are lower case Latin letters x, y, z, \dots . For example, if X is the number of children in a family, then x represents a specific integer 0, 1, 2, 3, Variables in statistics differ from variables in intermediate algebra in two following ways.

- The domain of the random variable (RV) is not necessarily a numerical set; the domain may be expressed in words; for example, if X = hair color then the domain is {black, blond, gray, green, orange}.
- We can tell what specific value x of the Random Variable X takes only after performing the experiment.

Comparing Two Independent Population Means with Known Population Standard Deviations

This module provides an overview of hypothesis testing in situations where there are both two independent population means and known population standard deviations in statistics.

Even though this situation is not likely (knowing the population standard deviations is not likely), the following example illustrates hypothesis testing for independent means, known population standard deviations. The sampling distribution for the difference between the means is normal and both populations must be normal. The random variable is $\bar{X} - \bar{X}$. The normal distribution has the following format:

Equation:

Normal distribution

$$\bar{X} - \bar{X} \sim N\left(u - u, \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}}\right)$$

Equation:

The standard deviation is:

$$\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}}$$

Equation:

The test statistic (z-score) is:

$$z = \frac{\bar{x} - \bar{x} - (\mu - \mu)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}}}$$

Example:

independent groups, population standard deviations known: The mean lasting time of 2 competing floor waxes is to be compared. **Twenty floors** are randomly assigned **to test each wax**. Both populations have a normal distribution. The following table is the result.

Wax	Sample Mean Number of Months Floor Wax Last	Population Standard Deviation
1	3	0.33
2	2.9	0.36

Exercise:

Problem:

Does the data indicate that **wax 1 is more effective than wax 2**? Test at a 5% level of significance.

Solution:

This is a test of two independent groups, two population means, population standard deviations known.

Random Variable: $\bar{X} - \bar{X}$ difference in the mean number of months the competing floor waxes last.

$$H_o: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

The words "is more effective" says that **wax 1 lasts longer than wax 2**, on the average. "Longer" is a $>$ symbol and goes into H_a . Therefore, this is a right-tailed test.

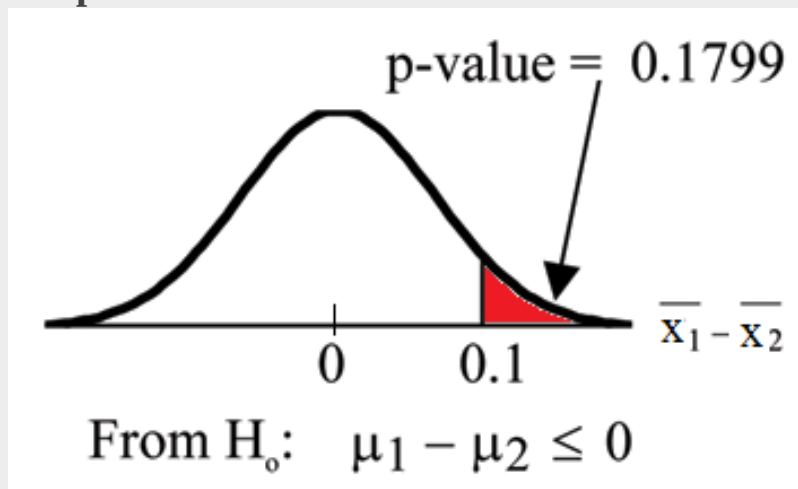
Distribution for the test: The population standard deviations are known so the distribution is normal. Using the formula above, the distribution is:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Since $\mu_1 > \mu_2$ then $\mu_1 - \mu_2 > 0$ and the mean for the normal distribution is 0.

Calculate the p-value using the normal distribution: p-value = 0.1799

Graph:



$$\alpha = 0.05$$

Compare α and the p-value: $\alpha = 0.05$ and p-value = 0.1799. Therefore, $\alpha >$ p-value.

Make a decision: Since $\alpha >$ p-value, do not reject H_o .

Conclusion: At the 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean time wax 1 lasts is longer (wax 1 is more effective) than the mean time wax 2 lasts.

Note: TI-83+ and TI-84: Press

STAT

. Arrow over to

TESTS

and press

3:2-SampZTest

. Arrow over to

Stats

and press

ENTER

. Arrow down and enter

.33

for sigma1,

.36

for sigma2,

3

for the first sample mean,

20

for n1,

2.9

for the second sample mean, and

20

for n2. Arrow down to μ_1 : and arrow to $> \mu_2$. Press

ENTER

. Arrow down to

Calculate

and press

ENTER

. The p-value is $p = 0.1799$ and the test statistic is 0.9157. Do the procedure again but instead of

Calculate

do

Draw

.

Comparing Two Independent Population Proportions

1. The two independent samples are simple random samples that are independent.
2. The number of successes is at least five and the number of failures is at least five for each of the samples.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance. A hypothesis test can help determine if a difference in the estimated proportions ($P'_A - P'_B$) reflects a difference in the population proportions.

The difference of two proportions follows an approximate normal distribution. Generally, the null hypothesis states that the two proportions are the same. That is, $H_o : p_A = p_B$. To conduct the test, we use a pooled proportion, p_c .

Equation:

The pooled proportion is calculated as follows:

$$p_c = \frac{x_A + x_B}{n_A + n_B}$$

Equation:

The distribution for the differences is:

$$P'_A - P'_B \sim N \left(0, \frac{p_c \cdot (1 - p_c)}{n_A} + \frac{p_c \cdot (1 - p_c)}{n_B} \right)$$

Equation:

The test statistic (z-score) is:

$$z = \frac{(p'_A - p'_B) - (p_A - p_B)}{\sqrt{p_c \cdot (1 - p_c) \cdot \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}}$$

Example:**Two population proportions**

Two types of medication for hives are being tested to determine if there is a **difference in the proportions of adult patient reactions**. **Twenty** out of a random **sample of 200** adults given medication A still had hives 30 minutes after taking the medication. **Twelve** out of another **random sample of 200 adults** given medication B still had hives 30 minutes after taking the medication. Test at a 1% level of significance.

Determining the solution

This is a test of 2 population proportions.

Exercise:

Problem: How do you know?

Solution:

The problem asks for a difference in proportions.

Let A and B be the subscripts for medication A and medication B. Then p_A and p_B are the desired population proportions.

Random Variable:

$P'_A - P'_B$ = difference in the proportions of adult patients who did not react after 30 minutes to medication A and medication B.

$$H_o : p_A = p_B \qquad p_A - p_B = 0$$

$$H_a : p_A \neq p_B \qquad p_A - p_B \neq 0$$

The words "**is a difference**" tell you the test is two-tailed.

Distribution for the test: Since this is a test of two binomial population proportions, the distribution is normal:

$$p_c = \frac{x_A + x_B}{n_A + n_B} = \frac{20 + 12}{200 + 200} = 0.08 \quad 1 - p_c = 0.92$$

Therefore,

$$P'_A - P'_B \sim N \left(0, \sqrt{0.08 \cdot 0.92 \cdot \left(\frac{1}{200} + \frac{1}{200} \right)} \right)$$

$P'_A - P'_B$ follows an approximate normal distribution.

Calculate the p-value using the normal distribution: p-value = 0.1404.

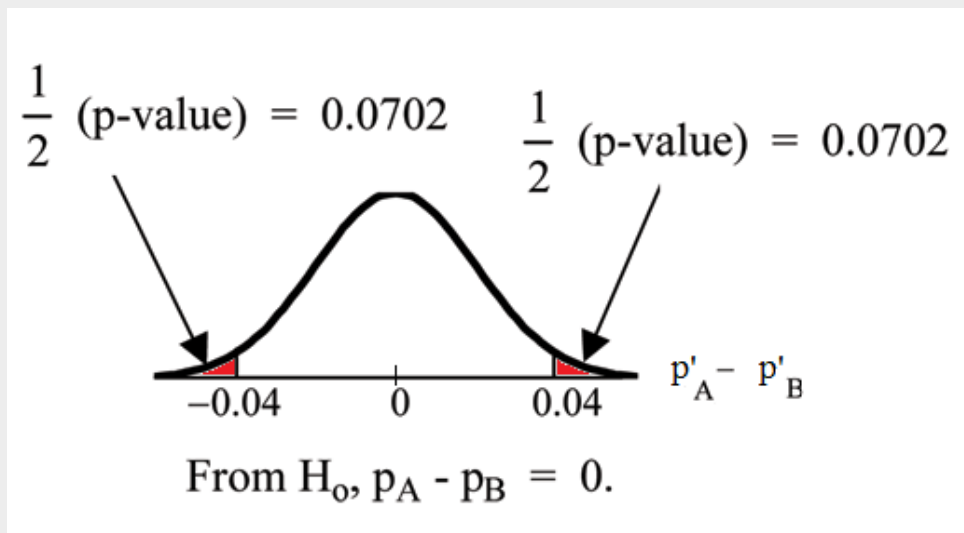
Estimated proportion for group A:

$$p'_A = \frac{x_A}{n_A} = \frac{20}{200} = 0.1$$

Estimated proportion for group B:

$$p'_B = \frac{x_B}{n_B} = \frac{12}{200} = 0.06$$

Graph:



$$P'_A - P'_B = 0.1 - 0.06 = 0.04.$$

Half the p-value is below -0.04 and half is above 0.04.

Compare α and the p-value: $\alpha = 0.01$ and the p-value = 0.1404. $\alpha < \text{p-value}$.

Make a decision: Since $\alpha < \text{p-value}$, do not reject H_o .

Conclusion: At a 1% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the proportions of adult patients who did not react after 30 minutes to medication A and medication B.

Note: TI-83+ and TI-84: Press

STAT

. Arrow over to

TESTS

and press

6:2-PropZTest

. Arrow down and enter

20

for x1,

200

for n1,

12

for x2, and

200

for n2. Arrow down to

p1

: and arrow to

not equal p2

. Press

ENTER

. Arrow down to

Calculate

and press

ENTER

. The p-value is $p = 0.1404$ and the test statistic is 1.47.
Do the procedure again but instead of

Calculate

do

Draw

.

Matched or Paired Samples

This module provides an overview of Hypothesis Testing: Matched or Paired Samples as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

1. Simple random sampling is used.
2. Sample sizes are often small.
3. Two measurements (samples) are drawn from the same pair of individuals or objects.
4. Differences are calculated from the matched or paired samples.
5. The differences form the sample that is used for the hypothesis test.
6. The matched pairs have differences that either come from a population that is normal or the number of differences is sufficiently large so the distribution of the sample mean of differences is approximately normal.

In a hypothesis test for matched or paired samples, subjects are matched in pairs and differences are calculated. The differences are the data. The population mean for the differences, μ_d , is then tested using a Student-t test for a single population mean with $n - 1$ degrees of freedom where n is the number of differences.

Equation:

The test statistic (t-score) is:

$$t = \frac{x_d - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

Example:

Matched or paired samples

A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in the table. The "before" value is matched to an "after" value and the differences are calculated. The differences have a normal distribution.

Subject:	A	B	C	D	E	F	G	H
Before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
After	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

Exercise:

Problem:

Are the sensory measurements, on average, lower after hypnotism? Test at a 5% significance level.

Solution:

Corresponding "before" and "after" values form matched pairs. (Calculate "after" - "before").

After Data	Before Data	Difference
6.8	6.6	0.2
2.4	6.5	-4.1
7.4	9	-1.6
8.5	10.3	-1.8
8.1	11.3	-3.2
6.1	8.1	-2
3.4	6.3	-2.9
2	11.6	-9.6

The data **for the test** are the differences: {0.2, -4.1, -1.6, -1.8, -3.2, -2, -2.9, -9.6}

The sample mean and sample standard deviation of the differences are: $\bar{x}_d = -3.13$ and $s_d = 2.91$
Verify these values.

Let μ_d be the population mean for the differences. We use the subscript d to denote "differences."

Random Variable: X_d = the mean difference of the sensory measurements

Equation:

$$H_o : \mu_d \geq 0$$

There is no improvement. (μ_d is the population mean of the differences.)

Equation:

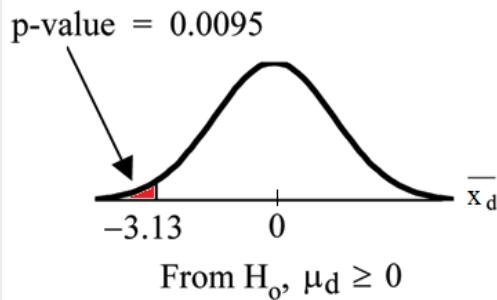
$$H_a : \mu_d < 0$$

There is improvement. The score should be lower after hypnotism so the difference ought to be negative to indicate improvement.

Distribution for the test: The distribution is a student-t with $df = n - 1 = 8 - 1 = 7$. Use t_7 . (**Notice that the test is for a single population mean.**)

Calculate the p-value using the Student-t distribution: p-value = 0.0095

Graph:



X_d is the random variable for the differences.

The sample mean and sample standard deviation of the differences are:

$$x_d = -3.13$$

$$s_d = 2.91$$

Compare α and the p-value: $\alpha = 0.05$ and p-value = 0.0095. $\alpha > \text{p-value}$.

Make a decision: Since $\alpha > \text{p-value}$, reject H_o .

This means that $\mu_d < 0$ and there is improvement.

Conclusion: At a 5% level of significance, from the sample data, there is sufficient evidence to conclude that the sensory measurements, on average, are lower after hypnotism. Hypnotism appears to be effective in reducing pain.

Note:For the TI-83+ and TI-84 calculators, you can either calculate the differences ahead of time (**after** - **before**) and put the differences into a list or you can put the **after** data into a first list and the **before** data into a second list. Then go to a third list and arrow up to the name. Enter 1st list name - 2nd list name. The calculator will do the subtraction and you will have the differences in the third list.

Note:TI-83+ and TI-84: Use your list of differences as the data. Press

STAT

and arrow over to

TESTS

. Press

2:T-Test

. Arrow over to

Data

and press

ENTER

. Arrow down and enter

0

for μ , the name of the list where you put the data, and

1

for Freq:. Arrow down to

μ

: and arrow over to

<

μ . Press

ENTER

. Arrow down to

Calculate

and press

ENTER

. The p-value is 0.0094 and the test statistic is -3.04. Do these instructions again except arrow to

Draw

(instead of

Calculate

). Press

ENTER

.

Example:

A college football coach was interested in whether the college's strength development class increased his players' maximum lift (in pounds) on the bench press exercise. He asked 4 of his players to participate in a study. The amount of weight they could each lift was recorded before they took the strength development class. After completing the class, the amount of weight they could each lift was again measured. The data are as follows:

Weight (in pounds)	Player 1	Player 2	Player 3	Player 4
Amount of weighted lifted prior to the class	205	241	338	368
Amount of weight lifted after the class	295	252	330	360

The coach wants to know if the strength development class makes his players stronger, on average.
Exercise:

Problem:

Record the **differences** data. Calculate the differences by subtracting the amount of weight lifted prior to the class from the weight lifted after completing the class. The data for the differences are: {90, 11, -8, -8}. The differences have a normal distribution.

Using the differences data, calculate the sample mean and the sample standard deviation.

$$x_d = 21.3 \quad s_d = 46.7$$

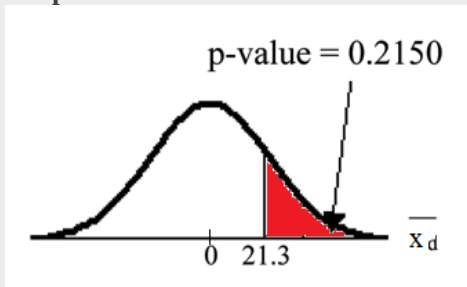
Using the difference data, this becomes a test of a single _____ (fill in the blank).

Define the random variable: X_d = mean difference in the maximum lift per player.

The distribution for the hypothesis test is t .

$$H_o : \mu_d \leq 0 \quad H_a : \mu_d > 0$$

Graph:



Calculate the p-value: The p-value is 0.2150

Decision: If the level of significance is 5%, the decision is to not reject the null hypothesis because $\alpha < \text{p-value}$.

What is the conclusion?

Solution:

means; At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the strength development class helped to make the players stronger, on average.

Example:

Seven eighth graders at Kennedy Middle School measured how far they could push the shot-put with their dominant (writing) hand and their weaker (non-writing) hand. They thought that they could push equal distances with either hand. The following data was collected.

Distance (in feet) using	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7
Dominant Hand	30	26	34	17	19	26	20
Weaker Hand	28	14	27	18	17	26	16

Exercise:

Problem:

Conduct a hypothesis test to determine whether the mean difference in distances between the children's dominant versus weaker hands is significant.

Note:use a t-test on the difference data. Assume the differences have a normal distribution. The random variable is the mean difference.

Note:The test statistic is 2.18 and the p-value is 0.0716.

What is your conclusion?

Solution:

H : μ_d equals 0; H_a : μ_d does not equal 0; Do not reject the null; At a 5% significance level, from the sample data, there is not sufficient evidence to conclude that the mean difference in distances between the children's dominant versus weaker hands is significant (there is not sufficient evidence to show that the children could push the shot-put further with their dominant hand). Alpha and the p-value are close so the test is not strong.

Summary of Types of Hypothesis Tests

Two Population Means

- Populations are independent and population standard deviations are unknown.
- Populations are independent and population standard deviations are known (not likely).

Matched or Paired Samples

- Two samples are drawn from the same set of objects.
- Samples are dependent.

Two Population Proportions

- Populations are independent.

Practice 1: Hypothesis Testing for Two Proportions

This module provides a practice of Two Population Means and Two Population Proportions as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

- The student will conduct a hypothesis test of two proportions.

Given

In the recent Census, 3 percent of the U.S. population reported being two or more races. However, the percent varies tremendously from state to state.

(Source: <http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf>)

Suppose that two random surveys are conducted. In the first random survey, out of 1000 North Dakotans, only 9 people reported being of two or more races. In the second random survey, out of 500 Nevadans, 17 people reported being of two or more races. Conduct a hypothesis test to determine if the population percents are the same for the two states or if the percent for Nevada is statistically higher than for North Dakota.

Hypothesis Testing: Two Proportions

Exercise:

Problem: Is this a test of means or proportions?

Solution:

Proportions

Exercise:

Problem: State the null and alternative hypotheses.

- a

- **b**

Solution:

- **a**
- **a**

Exercise:

Problem:

Is this a right-tailed, left-tailed, or two-tailed test? How do you know?

Solution:

right-tailed

Exercise:

Problem: What is the Random Variable of interest for this test?

Exercise:

Problem: In words, define the Random Variable for this test.

Exercise:

Problem:

Which distribution (Normal or student's-t) would you use for this hypothesis test?

Solution:

Normal

Exercise:

Problem:

Explain why you chose the distribution you did for the above question.

Exercise:

Problem: Calculate the test statistic.

Solution:

3.50

Exercise:**Problem:**

Sketch a graph of the situation. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the α value.

**Exercise:**

Problem: Find the α value:

Solution:

0.0002

Exercise:

Problem: At a pre-conceived α , what is your:

- **a** Decision:
 - **b** Reason for the decision:
 - **c** Conclusion (write out in a complete sentence):
-

Solution:

- **a** Reject the null hypothesis

Discussion Question

Exercise:

Problem:

Does it appear that the proportion of Nevadans who are two or more races is higher than the proportion of North Dakotans? Why or why not?

Practice 2: Hypothesis Testing for Two Averages

This module provides a practice of Hypothesis Testing: Two Population Means and Two Population Proportions: as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcome

- The student will conduct a hypothesis test of two means.

Given

The U.S. Center for Disease Control reports that the mean life expectancy for whites born in 1900 was 47.6 years and for nonwhites it was 33.0 years. (http://www.cdc.gov/nchs/data/dvs/nvsr53_06t12.pdf) Suppose that you randomly survey death records for people born in 1900 in a certain county. Of the 124 whites, the mean life span was 45.3 years with a standard deviation of 12.7 years. Of the 82 nonwhites, the mean life span was 34.1 years with a standard deviation of 15.6 years. Conduct a hypothesis test to see if the mean life spans in the county were the same for whites and nonwhites.

Hypothesis Testing: Two Means

Exercise:

Problem: Is this a test of means or proportions?

Solution:

Means

Exercise:

Problem: State the null and alternative hypotheses.

- a H_0 :

- **b** $H_a:$
-

Solution:

- **a** $H_0: \mu_W = \mu_{NW}$
- **b** $H_a: \mu_W \neq \mu_{NW}$

Exercise:

Problem:

Is this a right-tailed, left-tailed, or two-tailed test? How do you know?

Solution:

two-tailed

Exercise:

Problem: What is the Random Variable of interest for this test?

Solution:

$$X_W - X_{NW}$$

Exercise:

Problem:

In words, define the Random Variable of interest for this test.

Solution:

The difference between the mean life spans of whites and nonwhites.

Exercise:

Problem:

Which distribution (Normal or student's-t) would you use for this hypothesis test?

Exercise:**Problem:**

Explain why you chose the distribution you did for the above question.

Exercise:

Problem: Calculate the test statistic.

Solution:

5.42

Exercise:**Problem:**

Sketch a graph of the situation. Label the horizontal axis. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the p –value.

**Exercise:**

Problem: Find the p –value:

Solution:

0.0000

Exercise:

Problem: At a pre-conceived $\alpha = 0.05$, what is your:

- **a** Decision:
 - **b** Reason for the decision:
 - **c** Conclusion (write out in a complete sentence):
-

Solution:

- **a** Reject the null hypothesis

Discussion Question

Exercise:

Problem:

Does it appear that the means are the same? Why or why not?

Review

The next three questions refer to the following information:

In a survey at Kirkwood Ski Resort the following information was recorded:

	0 – 10	11 - 20	21 - 40	40+
Ski	10	12	30	8
Snowboard	6	17	12	5

Sport Participation by Age

Suppose that one person from of the above was randomly selected.

Exercise:

Problem:

Find the probability that the person was a skier or was age 11 – 20.

Solution:

$$\frac{77}{100}$$

Exercise:

Problem:

Find the probability that the person was a snowboarder given he/she was age 21 – 40.

Solution:

Exercise:

Problem: Explain which of the following are true and which are false.

- **a** Sport and Age are independent events.
 - **b** Ski and age 11 – 20 are mutually exclusive events.
 - **c** $(\text{Ski and age } 21 - 40) < (\text{Ski} \mid \text{age } 21 - 40)$
 - **d** $(\text{Snowboard or age } 0 - 10) < (\text{Snowboard} \mid \text{age } 0 - 10)$
-

Solution:

- **a** False
- **b** False
- **c** True
- **d** False

Exercise:**Problem:**

The average length of time a person with a broken leg wears a cast is approximately 6 weeks. The standard deviation is about 3 weeks. Thirty people who had recently healed from broken legs were interviewed. State the distribution that most accurately reflects total time to heal for the thirty people.

Solution:

(180 16.43)

Exercise:

Problem:

The distribution for X is Uniform. What can we say for certain about the distribution for Y when $a = 1$?

- **A** The distribution for Y is still Uniform with the same mean and standard dev. as the distribution for X .
 - **B** The distribution for Y is Normal with the different mean and a different standard deviation as the distribution for X .
 - **C** The distribution for Y is Normal with the same mean but a larger standard deviation than the distribution for X .
 - **D** The distribution for Y is Normal with the same mean but a smaller standard deviation than the distribution for X .
-

Solution:

A

Exercise:**Problem:**

The distribution for X is uniform. What can we say for certain about the distribution for Y when $a = 50$?

- **A** The distribution for Y is still uniform with the same mean and standard deviation as the distribution for X .
 - **B** The distribution for Y is Normal with the same mean but a larger standard deviation as the distribution for X .
 - **C** The distribution for Y is Normal with a larger mean and a larger standard deviation than the distribution for X .
 - **D** The distribution for Y is Normal with the same mean but a smaller standard deviation than the distribution for X .
-

Solution:

C

The next three questions refer to the following information:

A group of students measured the lengths of all the carrots in a five-pound bag of baby carrots. They calculated the average length of baby carrots to be 2.0 inches with a standard deviation of 0.25 inches. Suppose we randomly survey 16 five-pound bags of baby carrots.

Exercise:

Problem:

State the approximate distribution for \bar{x} , the distribution for the average lengths of baby carrots in 16 five-pound bags. $\bar{x} \sim$

Solution:

$$\left(2, \frac{.25}{\sqrt{16}}\right)$$

Exercise:

Problem:

Explain why we cannot find the probability that one individual randomly chosen carrot is greater than 2.25 inches.

Exercise:

Problem: Find the probability that \bar{x} is between 2 and 2.25 inches.

Solution:

0.5000

The next three questions refer to the following information:

At the beginning of the term, the amount of time a student waits in line at the campus store is normally distributed with a mean of 5 minutes and a standard deviation of 2 minutes.

Exercise:

Problem: Find the 90th percentile of waiting time in minutes.

Solution:

7.6

Exercise:

Problem: Find the median waiting time for one student.

Solution:

5

Exercise:

Problem:

Find the probability that the average waiting time for 40 students is at least 4.5 minutes.

Solution:

0.9431

Lab: Hypothesis Testing for Two Means and Two Proportions

Class Time:

Names:

Student Learning Outcomes:

- The student will select the appropriate distributions to use in each case.
- The student will conduct hypothesis tests and interpret the results.

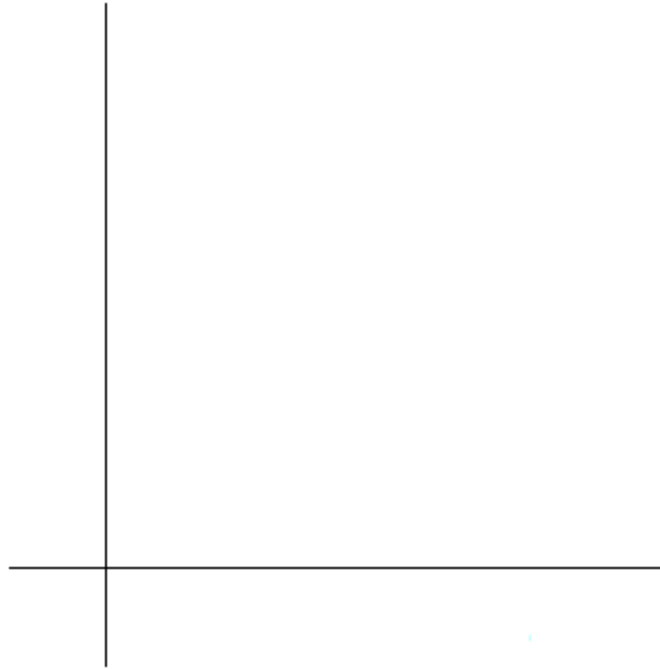
Supplies:

- The business section from two consecutive days' newspapers
- 3 small packages of M&Ms®
- 5 small packages of Reese's Pieces®

Increasing Stocks Survey

Look at yesterday's newspaper business section. Conduct a hypothesis test to determine if the proportion of New York Stock Exchange (NYSE) stocks that increased is greater than the proportion of NASDAQ stocks that increased. As randomly as possible, choose 40 NYSE stocks and 32 NASDAQ stocks and complete the following statements.

1. H_o
2. H_a
3. In words, define the Random Variable. _____ =
4. The distribution to use for the test is:
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.
 - aGraph:



- **b** Calculate the p-value:

7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Decreasing Stocks Survey

Randomly pick 8 stocks from the newspaper. Using two consecutive days' business sections, test whether the stocks went down, on average, for the second day.

1. H_o
2. H_a
3. In words, define the Random Variable. _____ =
4. The distribution to use for the test is:
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.

- **a**Graph:



- **b**Calculate the p-value:

7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

Candy Survey

Buy three small packages of M&Ms and 5 small packages of Reese's Pieces (same net weight as the M&Ms). Test whether or not the mean number of candy pieces per package is the same for the two brands.

1. H_o :
2. H_a :
3. In words, define the random variable. _____ =
4. What distribution should be used for this test?
5. Calculate the test statistic using your data.

6. Draw a graph and label it appropriately. Shade the actual level of significance.

◦ **a**Graph:



◦ **b**Calculate the p-value:

7. Do you reject or not reject the null hypothesis? Why?

8. Write a clear conclusion using a complete sentence.

Shoe Survey

Test whether women have, on average, more pairs of shoes than men. Include all forms of sneakers, shoes, sandals, and boots. Use your class as the sample.

1. H_o

2. H_a

3. In words, define the Random Variable. _____ =

4. The distribution to use for the test is:
5. Calculate the test statistic using your data.
6. Draw a graph and label it appropriately. Shade the actual level of significance.

◦ **a**Graph:



◦ **b**Calculate the p-value:

7. Do you reject or not reject the null hypothesis? Why?
8. Write a clear conclusion using a complete sentence.

The Chi-Square Distribution

This module provides an introduction to Chi-Square Distribution as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

By the end of this chapter, the student should be able to:

- Interpret the chi-square probability distribution as the sample size changes.
- Conduct and interpret chi-square goodness-of-fit hypothesis tests.
- Conduct and interpret chi-square test of independence hypothesis tests.
- Conduct and interpret chi-square homogeneity hypothesis tests.
- Conduct and interpret chi-square single variance hypothesis tests.

Introduction

Have you ever wondered if lottery numbers were evenly distributed or if some numbers occurred with a greater frequency? How about if the types of movies people preferred were different across different age groups? What about if a coffee machine was dispensing approximately the same amount of coffee each time? You could answer these questions by conducting a hypothesis test.

You will now study a new distribution, one that is used to determine the answers to the above examples. This distribution is called the Chi-square distribution.

In this chapter, you will learn the three major applications of the Chi-square distribution:

- The goodness-of-fit test, which determines if data fit a particular distribution, such as with the lottery example
- The test of independence, which determines if events are independent, such as with the movie example

- The test of a single variance, which tests variability, such as with the coffee example

Note: Though the Chi-square calculations depend on calculators or computers for most of the calculations, there is a table available (see the Table of Contents **15. Tables**). TI-83+ and TI-84 calculator instructions are included in the text.

Optional Collaborative Classroom Activity

Look in the sports section of a newspaper or on the Internet for some sports data (baseball averages, basketball scores, golf tournament scores, football odds, swimming times, etc.). Plot a histogram and a boxplot using your data. See if you can determine a probability distribution that your data fits. Have a discussion with the class about your choice.

Notation

This module provides an overview of Chi-Square Distribution Notation as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

The notation for the chi-square distribution is:

$$\chi^2 \sim \chi_{df}^2$$

where df = degrees of freedom depend on how chi-square is being used. (If you want to practice calculating chi-square probabilities then use $df = n - 1$. The degrees of freedom for the three major uses are each calculated differently.)

For the χ^2 distribution, the population mean is $\mu = df$ and the population standard deviation is $\sigma = \sqrt{2 \cdot df}$.

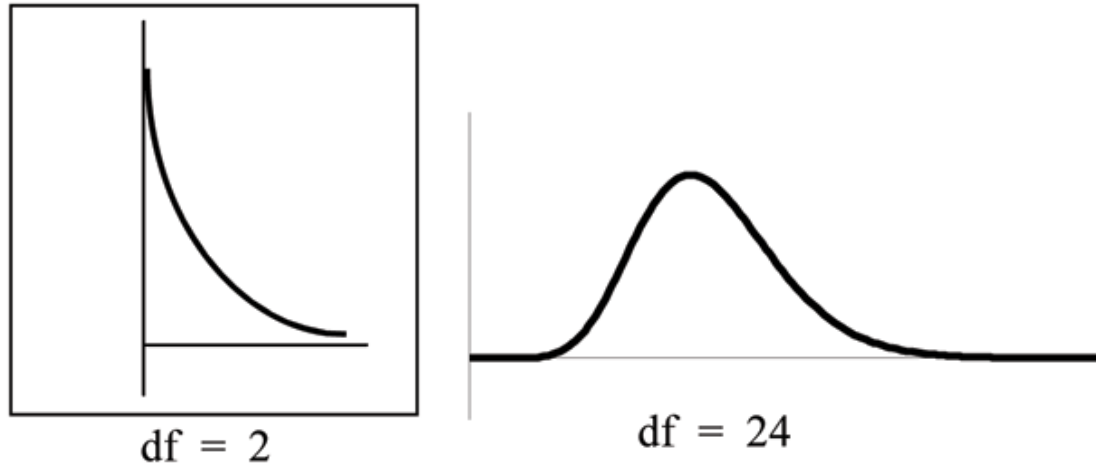
The random variable is shown as χ^2 but may be any upper case letter.

The random variable for a chi-square distribution with k degrees of freedom is the sum of k independent, squared standard normal variables.

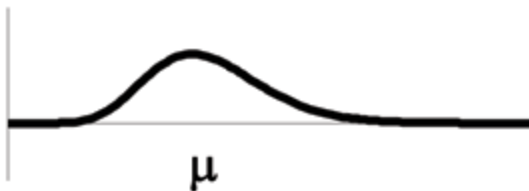
$$\chi^2 = Z_1^2 + Z_2^2 + \dots + Z_k^2$$

Facts About the Chi-Square Distribution

1. The curve is nonsymmetrical and skewed to the right.
2. There is a different chi-square curve for each df .



3. The test statistic for any test is always greater than or equal to zero.
4. When df is large, the chi-square curve approximates the normal. For $df \geq 30$, the mean, μ , and the standard deviation, σ , are approximately $\mu \approx df$ and $\sigma \approx \sqrt{2df}$. Therefore, $\chi^2 \sim N(df, \sqrt{2df})$, approximately.
5. The mean, μ , is located just to the right of the peak.



In the next sections, you will learn about four different applications of the Chi-Square Distribution. These hypothesis tests are almost always right-tailed tests. In order to understand why the tests are mostly right-tailed, you will need to look carefully at the actual definition of the test statistic. Think about the following while you study the next four sections. If the expected and observed values are "far" apart, then the test statistic will be "large" and we will reject in the right tail. The only way to obtain a test statistic very close to zero, would be if the observed and expected values are very, very close to each other. A left-tailed test could be used to determine if the fit were "too good." A "too good" fit might occur if data had been manipulated

or invented. Think about the implications of right-tailed versus left-tailed hypothesis tests as you learn the applications of the Chi-Square Distribution.

Goodness-of-Fit Test

This module describes how the chi-square distribution is used to conduct goodness-of-fit test.

In this type of hypothesis test, you determine whether the data "**fit**" a particular distribution or not. For example, you may suspect your unknown data fit a binomial distribution. You use a chi-square test (meaning the distribution for the hypothesis test is chi-square) to determine if there is a fit or not. **The null and the alternate hypotheses for this test may be written in sentences or may be stated as equations or inequalities.**

The test statistic for a goodness-of-fit test is:

Equation:

$$\sum_k \frac{(O - E)^2}{E}$$

where:

- O = observed values (data)
- E = expected values (from theory)
- k = the number of different data cells or categories

The observed values are the data values and the expected values are the values you would expect to get if the null hypothesis were true. There are n terms of the form $\frac{(O-E)^2}{E}$.

The degrees of freedom are $df = (\text{number of categories} - 1)$.

The goodness-of-fit test is almost always right tailed. If the observed values and the corresponding expected values are not close to each other, then the test statistic can get very large and will be way out in the right tail of the chi-square curve.

Note: The expected value for each cell needs to be at least 5 in order to use this test.

Example:

Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Suppose that a study was done to determine if the actual student absenteeism follows faculty perception. The

faculty expected that a group of 100 students would miss class according to the following chart.

Number absences per term	Expected number of students
0 - 2	50
3 - 5	30
6 - 8	12
9 - 11	6
12+	2

A random survey across all mathematics courses was then done to determine the actual number (**observed**) of absences in a course. The next chart displays the result of that survey.

Number absences per term	Actual number of students
0 - 2	35
3 - 5	40
6 - 8	20
9 - 11	1
12+	4

Determine the null and alternate hypotheses needed to conduct a goodness-of-fit test.
 H_o : Student absenteeism **fits** faculty perception.

The alternate hypothesis is the opposite of the null hypothesis.

H_a : Student absenteeism **does not fit** faculty perception.

Exercise:

Problem:

Can you use the information as it appears in the charts to conduct the goodness-of-fit test?

Solution:

No. Notice that the expected number of absences for the "12+" entry is less than 5 (it is 2). Combine that group with the "9 - 11" group to create new tables where the number of students for each entry are at least 5. The new tables are below.

Number absences per term	Expected number of students
0 - 2	50
3 - 5	30
6 - 8	12
9+	8

Number absences per term	Actual number of students
0 - 2	35
3 - 5	40
6 - 8	20
9+	5

Exercise:

Problem: What are the degrees of freedom (df)?

Solution:

There are 4 "cells" or categories in each of the new tables.

$$df = \text{number of cells} - 1 = 4 - 1 = 3$$

Example:

Employers particularly want to know which days of the week employees are absent in a five day work week. Most employers would like to believe that employees are absent equally during the week. Suppose a random sample of 60 managers were asked on which day of the week did they have the highest number of employee absences. The results were distributed as follows:

	Monday	Tuesday	Wednesday	Thursday	Friday
Number of Absences	15	12	9	9	15

Day of the Week Employees were most Absent

Exercise:**Problem:**

For the population of employees, do the days for the highest number of absences occur with equal frequencies during a five day work week? Test at a 5% significance level.

Solution:

The null and alternate hypotheses are:

- H_o : The absent days occur with equal frequencies, that is, they fit a uniform distribution.
- H_a : The absent days occur with unequal frequencies, that is, they do not fit a uniform distribution.

If the absent days occur with equal frequencies, then, out of 60 absent days (the total in the sample: $15 + 12 + 9 + 9 + 15 = 60$), there would be 12 absences on Monday, 12 on Tuesday, 12 on Wednesday, 12 on Thursday, and 12 on Friday. These numbers are the **expected** (E) values. The values in the table are the **observed** (O) values or data.

This time, calculate the χ^2 test statistic by hand. Make a chart with the following headings and fill in the columns:

- Expected (E) values (12, 12, 12, 12, 12)
- Observed (O) values (15, 12, 9, 9, 15)
- $(O - E)$
- $(O - E)^2$
- $\frac{(O-E)^2}{E}$

The last column ($\frac{(O-E)^2}{E}$) should have 0.75, 0, 0.75, 0.75, 0.75.

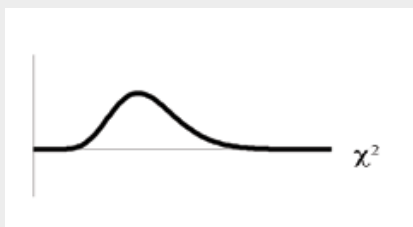
Now add (sum) the last column. Verify that the sum is 3. This is the χ^2 test statistic.

To find the p-value, calculate $P \chi^2 > 3$. This test is right-tailed. (Use a computer or calculator to find the p-value. You should get p-value = 0.5578.)

The dfs are the number of cells $- 1 = 5 - 1 = 4$.

TI-83+ and TI-84: Press **2nd DISTR**. Arrow down to χ^2 cdf. Press **ENTER**. Enter **(3, 10^99, 4)**. Rounded to 4 decimal places, you should see 0.5578 which is the p-value.

Next, complete a graph like the one below with the proper labeling and shading. (You should shade the right tail.)



The decision is to not reject the null hypothesis.

Conclusion: At a 5% level of significance, from the sample data, there is not sufficient evidence to conclude that the absent days do not occur with equal frequencies.

Note: TI-83+ and some TI-84 calculators do not have a special program for the test statistic for the goodness-of-fit test. The next example (Example 11-3) has the calculator instructions. The newer TI-84 calculators have in

STAT TESTS

the test

Chi2 GOF

. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press

STAT

TESTS

and

Chi2 GOF

. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press

calculate

or

draw

. Make sure you clear any lists before you start. See below.

Note: To Clear Lists in the calculators: Go into

STAT EDIT

and arrow up to the list name area of the particular list. Press

CLEAR

and then arrow down. The list will be cleared. Or, you can press

STAT

and press 4 (for

ClrList

). Enter the list name and press

ENTER

.

Example:

One study indicates that the number of televisions that American families have is distributed (this is the **given** distribution for the American population) as follows:

Number of Televisions	Percent
0	10
1	16
2	55
3	11
over 3	8

The table contains expected (E) percents.

A random sample of 600 families in the far western United States resulted in the following data:

Number of Televisions	Frequency
0	66
1	119
2	340
3	60
over 3	15
	Total = 600

The table contains observed (O) frequency values.

Exercise:

Problem:

At the 1% significance level, does it appear that the distribution "number of televisions" of far western United States families is different from the distribution for the American population as a whole?

Solution:

This problem asks you to test whether the far western United States families distribution fits the distribution of the American families. This test is always right-tailed.

The first table contains expected percentages. To get expected (E) frequencies, multiply the percentage by 600. The expected frequencies are:

Number of Televisions	Percent	Expected Frequency
0	10	$(0.10) \cdot (600) = 60$
1	16	$(0.16) \cdot (600) = 96$
2	55	$(0.55) \cdot (600) = 330$
3	11	$(0.11) \cdot (600) = 66$
over 3	8	$(0.08) \cdot (600) = 48$

Therefore, the expected frequencies are 60, 96, 330, 66, and 48. In the TI calculators, you can let the calculator do the math. For example, instead of 60, enter $.10 \cdot 600$.

H_o : The "number of televisions" distribution of far western United States families is the same as the "number of televisions" distribution of the American population.

H_a : The "number of televisions" distribution of far western United States families is different from the "number of televisions" distribution of the American population.

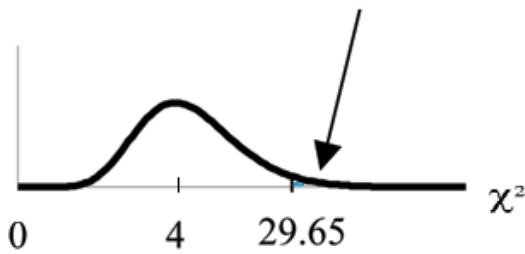
Distribution for the test: χ^2_4 where $df = (\text{the number of cells}) - 1 = 5 - 1 = 4$.

Note: $df \neq 600 - 1$

Calculate the test statistic: $\chi^2 = 29.65$

Graph:

p-value = 0.000006 (almost 0)



Probability statement: $p\text{-value} = P \chi^2 > 29.65 = 0.000006$.

Compare α and the p-value:

- $\alpha = 0.01$
- p-value = 0.000006

So, $\alpha > \text{p-value}$.

Make a decision: Since $\alpha > \text{p-value}$, reject H_o .

This means you reject the belief that the distribution for the far western states is the same as that of the American population as a whole.

Conclusion: At the 1% significance level, from the data, there is sufficient evidence to conclude that the "number of televisions" distribution for the far western United States is different from the "number of televisions" distribution for the American population as a whole.

Note: TI-83+ and some TI-84 calculators: Press

STAT

and

ENTER

. Make sure to clear lists

L1

,

L2

, and

L3

if they have data in them (see the note at the end of Example 11-2). Into

L1

, put the observed frequencies

66

,

119

,

349

,

60

,

15

. Into

L2

, put the expected frequencies

$.10 * 600$, $.16 * 600$

,

$.55 * 600$

,

$.11 * 600$

,

.08*600

. Arrow over to list

L3

and up to the name area

"L3"

. Enter

$(L1 - L2)^2 / L2$

and

ENTER

. Press

2nd QUIT

. Press

2nd LIST

and arrow over to

MATH

. Press

5

. You should see

"sum" (Enter L3)

. Rounded to 2 decimal places, you should see

29.65

. Press

2nd DISTR

. Press

7

or Arrow down to

7: χ^2 cdf

and press

ENTER

. Enter

(29 . 65 , 1E99 , 4)

. Rounded to 4 places, you should see

5 . 77E - 6 = . 000006

(rounded to 6 decimal places) which is the p-value.

The newer TI-84 calculators have in

STAT TESTS

the test

Chi2 GOF

. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press

STAT

TESTS

and

Chi2 GOF

. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press

calculate

or

draw

. Make sure you clear any lists before you start.

Example:

Exercise:

Problem:

Suppose you flip two coins 100 times. The results are 20 HH, 27 HT, 30 TH, and 23 TT. Are the coins fair? Test at a 5% significance level.

Solution:

This problem can be set up as a goodness-of-fit problem. The sample space for flipping two fair coins is {HH, HT, TH, TT}. Out of 100 flips, you would expect 25 HH, 25 HT, 25 TH, and 25 TT. This is the expected distribution. The question, "Are the coins fair?" is the same as saying, "Does the distribution of the coins (20 HH, 27 HT, 30 TH, 23 TT) fit the expected distribution?"

Random Variable: Let X = the number of heads in one flip of the two coins. X takes on the value 0, 1, 2. (There are 0, 1, or 2 heads in the flip of 2 coins.) Therefore, the **number of cells is 3**. Since X = the number of heads, the observed frequencies are 20 (for 2 heads), 57 (for 1 head), and 23 (for 0 heads or both tails). The expected frequencies are 25 (for 2 heads), 50 (for 1 head), and 25 (for 0 heads or both tails). This test is right-tailed.

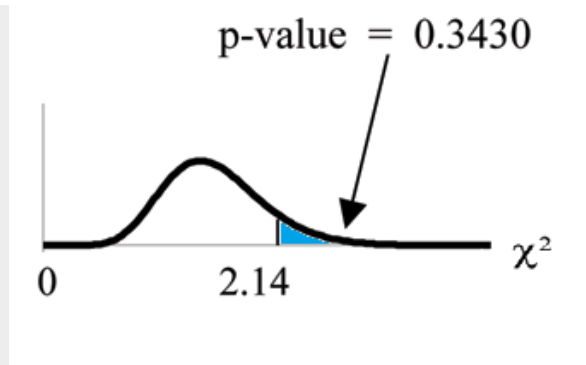
H_o : The coins are fair.

H_a : The coins are not fair.

Distribution for the test: χ^2_2 where $df = 3 - 1 = 2$.

Calculate the test statistic: $\chi^2 = 2.14$

Graph:



Probability statement: $p\text{-value} = P \chi^2 > 2.14 = 0.3430$

Compare α and the p-value:

- $\alpha = 0.05$
- $p\text{-value} = 0.3430$

So, $\alpha < p\text{-value}$.

Make a decision: Since $\alpha < p\text{-value}$, do not reject H_o .

Conclusion: There is insufficient evidence to conclude that the coins are not fair.

Note: TI-83+ and some TI- 84 calculators: Press

STAT

and

ENTER

. Make sure you clear lists

L1

,

L2

, and

L3

if they have data in them. Into

L1

, put the observed frequencies

20

,

57

,

23

. Into

L2

, put the expected frequencies

25

,

50

,

25

. Arrow over to list

L3

and up to the name area

"L3"

. Enter

$(L1 - L2)^2 / L2$

and

ENTER

. Press

2nd QUIT

. Press

2nd LIST

and arrow over to

MATH

. Press

5

. You should see

"sum"

.

Enter L3

. Rounded to 2 decimal places, you should see

2.14

. Press

2nd DISTR

. Arrow down to

7: χ^2 cdf

(or press

7

). Press

ENTER

. Enter

2.14, 1E99, 2)

. Rounded to 4 places, you should see

.3430

which is the p-value.

The newer TI-84 calculators have in

STAT TESTS

the test

Chi2 GOF

. To run the test, put the observed values (the data) into a first list and the expected values (the values you expect if the null hypothesis is true) into a second list. Press

STAT

TESTS

and

Chi2 GOF

. Enter the list names for the Observed list and the Expected list. Enter the degrees of freedom and press

calculate

or

draw

. Make sure you clear any lists before you start.

Test of Independence

This module describes how the chi-square distribution can be used to test for independence.

Tests of independence involve using a [contingency table](#) of observed (data) values. You first saw a contingency table when you studied probability in the [Probability Topics](#) chapter.

The test statistic for a test of independence is similar to that of a goodness-of-fit test:

Equation:

$$\sum_{i,j} \frac{O_{ij} - E_{ij}}{E_{ij}}$$

where:

- O = observed values
- E = expected values
- i = the number of rows in the table
- j = the number of columns in the table

There are $i \cdot j$ terms of the form $\frac{O_{ij} - E_{ij}}{E_{ij}}$.

A test of independence determines whether two factors are independent or not. You first encountered the term independence in Chapter 3. As a review, consider the following example.

Note: The expected value for each cell needs to be at least 5 in order to use this test.

Example:

Suppose A = a speeding violation in the last year and B = a cell phone user while driving. If A and B are independent then $P(A \cap B) = P(A)P(B)$. $A \cap B$ is the event that a driver received a speeding violation last year and is also a cell phone user while driving.

Suppose, in a study of drivers who received speeding violations in the last year and who uses cell phones while driving, that 755 people were surveyed. Out of the 755, 70 had a speeding violation and 685 did not; 305 were cell phone users while driving and 450 were not.

Let y = expected number of drivers that use a cell phone while driving and received speeding violations.

If A and B are independent, then $P(A \cap B) = P(A)P(B)$. By substitution,
 $\frac{y}{755} = \frac{70}{755} \cdot \frac{305}{755}$

Solve for y : $y = \frac{70 \cdot 305}{755}$

About 28 people from the sample are expected to be cell phone users while driving and to receive speeding violations.

In a test of independence, we state the null and alternate hypotheses in words. Since the contingency table consists of **two factors**, the null hypothesis states that the factors are **independent** and the alternate hypothesis states that they are **not independent (dependent)**.

If we do a test of independence using the example above, then the null hypothesis is:

H_0 : Being a cell phone user while driving and receiving a speeding violation are independent events.

If the null hypothesis were true, we would expect about 28 people to be cell phone users while driving and to receive a speeding violation.

The test of independence is always right-tailed because of the calculation of the test statistic. If the expected and observed values are not close together, then the test statistic is very large and way out in the right tail of the chi-square curve, like goodness-of-fit.

The degrees of freedom for the test of independence are:

The following formula calculates the **expected number (E)**:

E _____

Example:

In a volunteer group, adults 21 and older volunteer from one to nine hours each week to spend time with a disabled senior citizen. The program recruits among community college students, four-year college students, and nonstudents. The following table is a **sample** of the adult volunteers and the number of hours they volunteer per week.

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours	Row Total
Community College Students	111	96	48	255
Four-Year College Students	96	133	61	290
Nonstudents	91	150	53	294
Column Total	298	379	162	839

Number of Hours Worked Per Week by Volunteer Type (Observed)The table contains **observed (O)** values (data).

Exercise:

Problem: Are the number of hours volunteered **independent** of the type of volunteer?

Solution:

The **observed table** and the question at the end of the problem, "Are the number of hours volunteered independent of the type of volunteer?" tell you this is a test of independence. The two factors are **number of hours volunteered** and **type of volunteer**. This test is always right-tailed.

H_o : The number of hours volunteered is **independent** of the type of volunteer.

H_a : The number of hours volunteered is **dependent** on the type of volunteer.

The expected table is:

Type of Volunteer	1-3 Hours	4-6 Hours	7-9 Hours
Community College Students	90.57	115.19	49.24
Four-Year College Students	103.00	131.00	56.00
Nonstudents	104.42	132.81	56.77

Number of Hours Worked Per Week by Volunteer Type (Expected)The table contains **expected (E)** values (data).

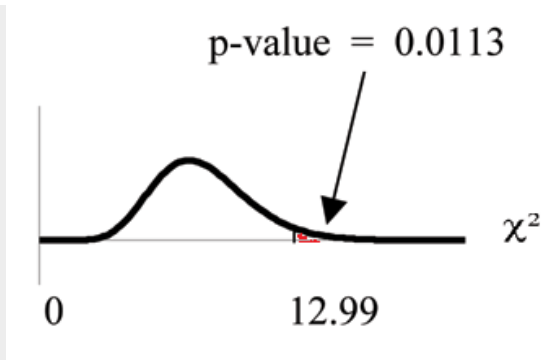
For example, the calculation for the expected frequency for the top left cell is

E _____

Calculate the test statistic: χ _____ (calculator or computer)

Distribution for the test: χ

Graph:



Probability statement: $P \chi^2$

Compare α and the : Since no α is given, assume α .
 α .

Make a decision: Since α , reject H_o . This means that the factors are not independent.

Conclusion: At a 5% level of significance, from the data, there is sufficient evidence to conclude that the number of hours volunteered and the type of volunteer are dependent on one another.

For the above example, if there had been another type of volunteer, teenagers, what would the degrees of freedom be?

Note: Calculator instructions follow.

TI-83+ and TI-84 calculator: Press the **MATRIX** key and arrow over to **EDIT**. Press **1: [A]**. Press **3 ENTER 3 ENTER**. Enter the table values by row from Example 11-6. Press **ENTER** after each. Press **2nd QUIT**. Press **STAT** and arrow over to **TESTS**. Arrow down to **C: χ^2 -TEST**. Press **ENTER**. You should see **Observed: [A]** and **Expected: [B]**. Arrow down to **Calculate**. Press **ENTER**. The test statistic is 12.9909 and the . Do the procedure a second time but arrow down to **Draw** instead of **calculate**.

Example:

De Anza College is interested in the relationship between anxiety level and the need to succeed in school. A random sample of 400 students took a test that measured anxiety level and need to succeed in school. The table shows the results. De Anza College wants to know if anxiety level and need to succeed in school are independent events.

Need to Succeed in School	High Anxiety	Med-high Anxiety	Medium Anxiety	Med-low Anxiety	Low Anxiety	Row Total
High Need	35	42	53	15	10	155
Medium Need	18	48	63	33	31	193
Low Need	4	5	11	15	17	52
Column Total	57	95	127	63	58	400

Need to Succeed in School vs. Anxiety Level

Exercise:

Problem:

How many high anxiety level students are expected to have a high need to succeed in school?

Solution:

The column total for a high anxiety level is 57. The row total for high need to succeed in school is 155. The sample size or total surveyed is 400.

E _____

The expected number of students who have a high anxiety level and a high need to succeed in school is about 22.

Exercise:

Problem:

If the two variables are independent, how many students do you expect to have a low need to succeed in school and a med-low level of anxiety?

Solution:

The column total for a med-low anxiety level is 63. The row total for a low need to succeed in school is 52. The sample size or total surveyed is 400.

Exercise:**Problem:**

- **a** E _____ =
- **b** The expected number of students who have a med-low anxiety level and a low need to succeed in school is about:

Solution:

- **a** E _____
- **b** 8

Glossary**Contingency Table**

The method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other. The table provides an easy way to calculate conditional probabilities.

Summary of Formulas

This module provides a summary on formulas used in Chi-Square Distribution as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

The Chi-Square Probability Distribution

$$\mu = df \text{ and } \sigma = \sqrt{2 \cdot df}$$

Goodness-of-Fit Hypothesis Test

- Use goodness-of-fit to test whether a data set fits a particular probability distribution.
- The degrees of freedom are number of cells or categories - 1.
- The test statistic is $\sum_k \frac{(O-E)^2}{E}$, where O = observed values (data), E = expected values (from theory), and k = the number of different data cells or categories.
- The test is right-tailed.

Test of Independence

- Use the test of independence to test whether two factors are independent or not.
- The degrees of freedom are equal to (number of columns - 1)(number of rows - 1).
- The test statistic is $\sum_{(i,j)} \frac{(O-E)^2}{E}$ where O = observed values, E = expected values, i = the number of rows in the table, and j = the number of columns in the table.
- The test is right-tailed.
- If the null hypothesis is true, the expected number
$$E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}}.$$

Test of Homogeneity

- Use the test for homogeneity to decide if two populations with unknown distributions have the same distribution as each other.
- The degrees of freedom are equal to number of columns - 1.

- The test statistic is $\sum_{(i,j)} \frac{(O-E)^2}{E}$ where O = observed values, E = expected values, i = the number of rows in the table, and j = the number of columns in the table.
- The test is right-tailed.
- If the null hypothesis is true, the expected number $E = \frac{(\text{row total})(\text{column total})}{\text{total surveyed}}$.

Note: The expected value for each cell needs to be at least 5 in order to use the Goodness-of-Fit, Independence and Homogeneity tests.

Test of a Single Variance

- Use the test to determine variation.
- The degrees of freedom are the number of samples - 1.
- The test statistic is $\frac{(n-1) \cdot s^2}{\sigma^2}$, where n = the total number of data, s^2 = sample variance, and σ^2 = population variance.
- The test may be left, right, or two-tailed.

Practice 1: Goodness-of-Fit Test

This module provides a practice on Chi-Square Distribution as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

- The student will conduct a goodness-of-fit test.

Given

The following data are real. The cumulative number of AIDS cases reported for Santa Clara County is broken down by ethnicity as follows: (Source: *HIV/AIDS Epidemiology Santa Clara County, Santa Clara County Public Health Department, May 2011*)

Ethnicity	Number of Cases
White	2229
Hispanic	1157
Black/African-American	457
Asian, Pacific Islander	232
	Total = 4075

The percentage of each ethnic group in Santa Clara County is as follows:

Ethnicity	Percentage of total county population	Number expected (round to 2 decimal places)
White	42.9%	1748.18
Hispanic	26.7%	
Black/African-American	2.6%	
Asian, Pacific Islander	27.8%	
	Total = 100%	

Expected Results

If the ethnicity of AIDS victims followed the ethnicity of the total county population, fill in the expected number of cases per ethnic group.

Goodness-of-Fit Test

Perform a goodness-of-fit test to determine whether the make-up of AIDS cases follows the ethnicity of the general population of Santa Clara County.

Exercise:

Problem: H_o :

Exercise:

Problem: H_a :

Exercise:

Problem: Is this a right-tailed, left-tailed, or two-tailed test?

Exercise:

Problem: degrees of freedom =

Solution:

degrees of freedom = 3

Exercise:

Problem: test statistic =

Solution:

2016.14

Exercise:

Problem: p-value =

Solution:

Rounded to 4 decimal places, the p-value is 0.0000.

Exercise:

Problem:

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.



Let α

Decision:

Reason for the Decision:

Conclusion (write out in complete sentences):

Discussion Question

Exercise:

Problem:

Does it appear that the pattern of AIDS cases in Santa Clara County corresponds to the distribution of ethnic groups in this county? Why or why not?

Practice 2: Contingency Tables

This module provides a practice on Chi-Square Distribution as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Student Learning Outcomes

- The student will conduct a test for independence using contingency tables.

Conduct a hypothesis test to determine if smoking level and ethnicity are independent.

Collect the Data

Copy the data provided in **Probability Topics Practice 1: Contingency Tables** into the table below.

Smoking Level Per Day	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1-10						
11-20						
21-30						
31+						
TOTALS						

Smoking Levels by Ethnicity (Observed)

Hypothesis

State the hypotheses.

- H_o :
- H_a :

Expected Values

Enter expected values in the above below. Round to two decimal places.

Analyze the Data

Calculate the following values:

Exercise:

Problem: Degrees of freedom =

Solution:

12

Exercise:

Problem: χ^2 test statistic =

Solution:

10301.8

Exercise:

Problem: p-value =

Solution:

0

Exercise:

Problem: Is this a right-tailed, left-tailed, or two-tailed test? Explain why.

Solution:

right

Graph the Data**Exercise:**

Problem:

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade in the region corresponding to the p-value.

**Conclusions**

State the decision and conclusion (in a complete sentence) for the following preconceived levels of α .
Exercise:

Problem: $\alpha = 0.05$

- **a** Decision:
 - **b** Reason for the decision:
 - **c** Conclusion (write out in a complete sentence):
-

Solution:

- **a** Reject the null hypothesis

Exercise:

Problem: $\alpha = 0.01$

- **a** Decision:
- **b** Reason for the decision:
- **c** Conclusion (write out in a complete sentence):

Practice 3: Test of a Single Variance

This module provides a practice on Chi-Square Distribution as a part of Elementary Statistics textbook.

Student Learning Outcomes

- The student will conduct a test of a single variance.

Given

Suppose an airline claims that its flights are consistently on time with an average delay of at most 15 minutes. It claims that the average delay is so consistent that the variance is no more than 150 minutes. Doubting the consistency part of the claim, a disgruntled traveler calculates the delays for his next 25 flights. The average delay for those 25 flights is 22 minutes with a standard deviation of 15 minutes.

Sample Variance

Exercise:

Problem:

Is the traveler disputing the claim about the average or about the variance?

Exercise:

Problem:

A sample standard deviation of 15 minutes is the same as a sample variance of _____ minutes.

Solution:

225

Exercise:

Problem: Is this a right-tailed, left-tailed, or two-tailed test?

Hypothesis Test

Perform a hypothesis test on the consistency part of the claim.

Exercise:

Problem: H_o :

Exercise:

Problem: H_a :

Exercise:

Problem: Degrees of freedom =

Solution:

24

Exercise:

Problem: Chi^2 test statistic =

Solution:

36

Exercise:

Problem: p-value =

Solution:

0.0549

Exercise:

Problem:

Graph the situation. Label and scale the horizontal axis. Mark the mean and test statistic. Shade the p-value.



Exercise:

Problem: Let $\alpha = 0.05$

Decision:

Conclusion (write out in a complete sentence):

Discussion Questions

Exercise:

Problem: How did you know to test the variance instead of the mean?

Exercise:

Problem:

If an additional test were done on the claim of the average delay, which distribution would you use?

Exercise:

Problem:

If an additional test was done on the claim of the average delay, but 45 flights were surveyed, which distribution would you use?

Review

This module provides an review on Chi-Square Distribution as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

The next two questions refer to the following real study:

A recent survey of U.S. teenage pregnancy was answered by 720 girls, age 12 - 19. 6% of the girls surveyed said they have been pregnant. (*Parade Magazine*) We are interested in the true proportion of U.S. girls, age 12 - 19, who have been pregnant.

Exercise:

Problem:

Find the 95% confidence interval for the true proportion of U.S. girls, age 12 - 19, who have been pregnant.

Solution:

(0.0424,0.0770)

Exercise:

Problem:

The report also stated that the results of the survey are accurate to within $\pm 3.7\%$ at the 95% confidence level. Suppose that a new study is to be done. It is desired to be accurate to within 2% of the 95% confidence level. What is the minimum number that should be surveyed?

Solution:

2401

Exercise:

Problem:

Given: $X \sim \text{Exp } \frac{1}{3}$. Sketch the graph that depicts: $(X > 1)$.

The next four questions refer to the following information:

Suppose that the time that owners keep their cars (purchased new) is normally distributed with a mean of 7 years and a standard deviation of 2 years. We are interested in how long an individual keeps his car (purchased new). Our population is people who buy their cars new.

Exercise:

Problem:

60% of individuals keep their cars **at most** how many years?

Solution:

7.5

Exercise:

Problem:

Suppose that we randomly survey one person. Find the probability that person keeps his/her car **less than** 2.5 years.

Solution:

0.0122

Exercise:

Problem:

If we are to pick individuals 10 at a time, find the distribution for the **mean** car length ownership.

Solution:

(7,0.63)

Exercise:

Problem:

If we are to pick 10 individuals, find the probability that the **sum** of their ownership time is more than 55 years.

Solution:

0.9911

Exercise:

Problem: For which distribution is the median not equal to the mean?

- AUniform
- BExponential
- CNormal
- DStudent-t

Solution:

B

Exercise:**Problem:**

Compare the standard normal distribution to the student-t distribution, centered at 0. Explain which of the following are true and which are false.

- **a**As the number surveyed increases, the area to the left of -1 for the student-t distribution approaches the area for the standard normal distribution.
- **b**As the degrees of freedom decrease, the graph of the student-t distribution looks more like the graph of the standard normal distribution.
- **c**If the number surveyed is 15, the normal distribution should never be used.

Solution:

- **a**True
- **b**False
- **c**False

The next five questions refer to the following information:

We are interested in the checking account balance of a twenty-year-old college student. We randomly survey 16 twenty-year-old college students. We obtain a sample mean of \$640 and a sample standard deviation of \$150. Let X = checking account balance of an individual twenty year old college student.

Exercise:

Problem: Explain why we cannot determine the distribution of X .

Exercise:**Problem:**

If you were to create a confidence interval or perform a hypothesis test for the population mean checking account balance of 20-year old college students, what distribution would you use?

Solution:

student-t with $df = 15$

Exercise:**Problem:**

Find the 95% confidence interval for the true mean checking account balance of a twenty-year-old college student.

Solution:

(560.07,719.93)

Exercise:

Problem:

What type of data is the balance of the checking account considered to be?

Solution:

quantitative - continuous

Exercise:

Problem:

What type of data is the number of 20 year olds considered to be?

Solution:

quantitative - discrete

Exercise:

Problem:

On average, a busy emergency room gets a patient with a shotgun wound about once per week. We are interested in the number of patients with a shotgun wound the emergency room gets per 28 days.

- **a** Define the random variable .
 - **b** State the distribution for .
 - **c** Find the probability that the emergency room gets no patients with shotgun wounds in the next 28 days.
-

Solution:

- **b** (4)
- **c** 0.0183

The next two questions refer to the following information:

The probability that a certain slot machine will pay back money when a quarter is inserted is 0.30 . Assume that each play of the slot machine is independent from each other. A person puts in 15 quarters for 15 plays.

Exercise:

Problem:

Is the expected number of plays of the slot machine that will pay back money greater than, less than or the same as the median? Explain your answer.

Solution:

greater than

Exercise:

Problem:

Is it likely that exactly 8 of the 15 plays would pay back money? Justify your answer numerically.

Solution:

No; $(\quad = 8) = 0.0348$

Exercise:

Problem: A game is played with the following rules:

- it costs \$10 to enter
- a fair coin is tossed 4 times
- if you do not get 4 heads or 4 tails, you lose your \$10
- if you get 4 heads or 4 tails, you get back your \$10, plus \$30 more

Over the long run of playing this game, what are your expected earnings?

Solution:

You will lose \$5

Exercise:**Problem:**

- The mean grade on a math exam in Rachel's class was 74, with a standard deviation of 5. Rachel earned an 80.
- The mean grade on a math exam in Becca's class was 47, with a standard deviation of 2. Becca earned a 51.
- The mean grade on a math exam in Matt's class was 70, with a standard deviation of 8. Matt earned an 83.

Find whose score was the best, compared to his or her own class.
Justify your answer numerically.

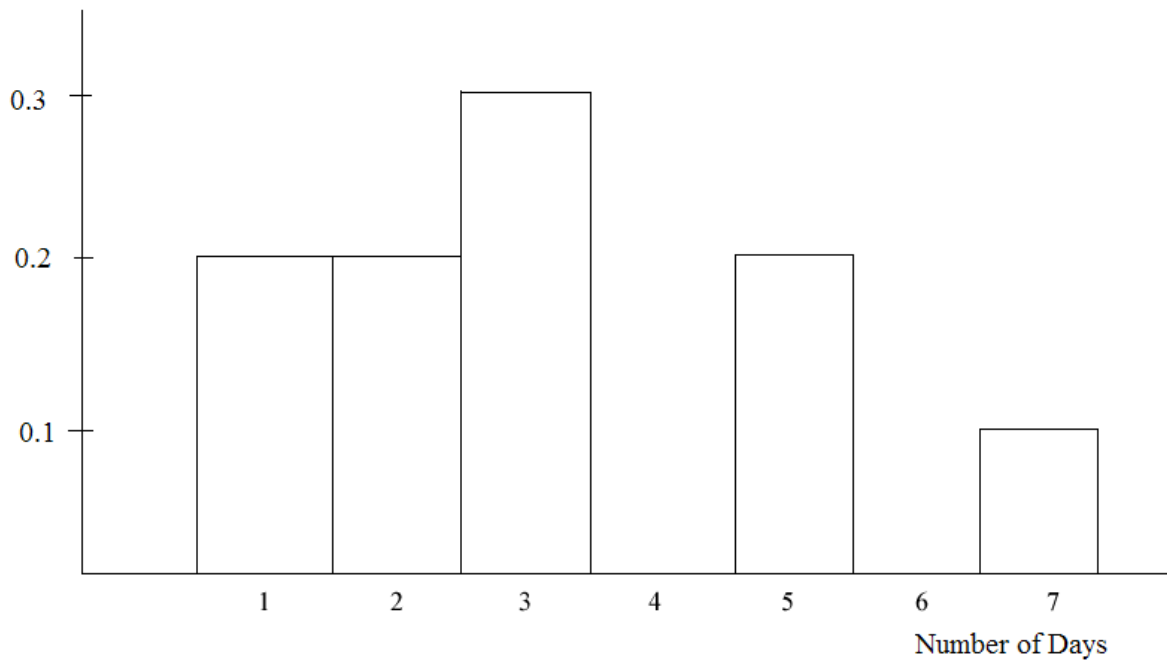
Solution:

Becca

The next two questions refer to the following information:

A random sample of 70 compulsive gamblers were asked the number of days they go to casinos per week. The results are given in the following graph:

Relative Frequency



Exercise:

Problem: Find the number of responses that were “5”.

Solution:

14

Exercise:

Problem:

Find the mean, standard deviation, the median, the first quartile, the third quartile and the IQR.

Solution:

- Sample mean = 3.2
- Sample standard deviation = 1.85
- Median = 3
- Quartile 1 = 2

- Quartile 3 = 5
- IQR = 3

Exercise:

Problem:

Based upon research at De Anza College, it is believed that about 19% of the student population speaks a language other than English at home.

Suppose that a study was done this year to see if that percent has decreased. Ninety-eight students were randomly surveyed with the following results. Fourteen said that they speak a language other than English at home.

- **a** State an appropriate **null** hypothesis.
- **b** State an appropriate **alternate** hypothesis.
- **c** Define the Random Variable, \bar{p} .
- **d** Calculate the test statistic.
- **e** Calculate the p-value.
- **f** At the 5% level of decision, what is your decision about the null hypothesis?
- **g** What is the Type I error?
- **h** What is the Type II error?

Solution:

- **d** $= -1.19$
- **e** 0.1171
- **f** Do not reject the null

Exercise:

Problem:

Assume that you are an emergency paramedic called in to rescue victims of an accident. You need to help a patient who is bleeding profusely. The patient is also considered to be a high risk for contracting AIDS. Assume that the null hypothesis is that the patient does **not** have the HIV virus. What is a Type I error?

Solution:

We conclude that the patient does have the HIV virus when, in fact, the patient does not.

Exercise:**Problem:**

It is often said that Californians are more casual than the rest of Americans. Suppose that a survey was done to see if the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals. Fifty of each was surveyed with the following results. 15 Californians wear jeans to work and 6 non-Californians wear jeans to work.

- C = Californian professional
 - NC = non-Californian professional
 - a) State appropriate **null** and **alternate** hypotheses.
 - b) Define the Random Variable.
 - c) Calculate the test statistic and p-value.
 - d) At the 5% significance level, what is your decision?
 - e) What is the Type I error?
 - f) What is the Type II error?
-

Solution:

- c) $z = 2.21$; $p = 0.0136$
- d) Reject the null

- **e**We conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is not greater.
- **f**We cannot conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is greater.

The next two questions refer to the following information:

A group of Statistics students have developed a technique that they feel will lower their anxiety level on statistics exams. They measured their anxiety level at the start of the quarter and again at the end of the quarter. Recorded is the paired data in that order: (1000, 900); (1200, 1050); (600, 700); (1300, 1100); (1000, 900); (900, 900).

Exercise:

Problem: This is a test of (pick the best answer):

- **A**large samples, independent means
- **B**small samples, independent means
- **C**dependent means

Solution:

C

Exercise:

Problem: State the distribution to use for the test.

Solution:

Practice Final Exam 1

This module is a practice final for an associated elementary statistics textbook, Collaborative Statistics.

Questions 1-2 refer to the following:

An experiment consists of tossing two 12-sided dice (the numbers 1-12 are printed on the sides of each dice).

- Let Event A = both dice show an even number
- Let Event B = both dice show a number more than 8

Exercise:

Problem: Events A and B are:

- A Mutually exclusive.
- B Independent.
- C Mutually exclusive and independent.
- D Neither mutually exclusive nor independent.

Solution:

B: Independent.

Exercise:

Problem: Find $P(A|B)$

- A $\frac{2}{4}$
- B $\frac{16}{144}$
- C $\frac{4}{16}$
- D $\frac{2}{144}$

Solution:

C: $\frac{4}{16}$

Exercise:

Problem:

Which of the following are TRUE when we perform a hypothesis test on matched or paired samples?

- A Sample sizes are almost never small.
- B Two measurements are drawn from the same pair of individuals or objects.
- C Two sample means are compared to each other.
- D Answer choices B and C are both true.

Solution:

B: Two measurements are drawn from the same pair of individuals or objects.

Questions 4 - 5 refer to the following:

118 students were asked what type of color their bedrooms were painted: light colors, dark colors or vibrant colors. The results were tabulated according to gender.

	Light colors	Dark colors	Vibrant colors
Female	20	22	28
Male	10	30	8

Exercise:

Problem:

Find the probability that a randomly chosen student is male or has a bedroom painted with light colors.

- A $\frac{10}{118}$
- B $\frac{68}{118}$
- C $\frac{48}{118}$
- D $\frac{10}{48}$

Solution:

B: $\frac{68}{118}$

Exercise:

Problem:

Find the probability that a randomly chosen student is male given the student's bedroom is painted with dark colors.

- A $\frac{30}{118}$
- B $\frac{30}{48}$
- C $\frac{22}{118}$
- D $\frac{30}{52}$

Solution:

D: $\frac{30}{52}$

Questions 6 – 7 refer to the following:

We are interested in the number of times a teenager must be reminded to do his/her chores each week. A survey of 40 mothers was conducted. The table below shows the results of the survey.

x	$P(x)$
0	$\frac{2}{40}$
1	$\frac{5}{40}$

2	
3	$\frac{14}{40}$
4	$\frac{7}{40}$
5	$\frac{4}{40}$

Exercise:

Problem: Find the probability that a teenager is reminded 2 times.

- A8
- B $\frac{8}{40}$
- C $\frac{6}{40}$
- D2

Solution:

B: $\frac{8}{40}$

Exercise:

Problem: Find the expected number of times a teenager is reminded to do his/her chores.

- A15
- B2.78
- C1.0
- D3.13

Solution:

B: 2.78

Questions 8 – 9 refer to the following:

On any given day, approximately 37.5% of the cars parked in the De Anza parking structure are parked crookedly. (Survey done by Kathy Plum.) We randomly survey 22 cars. We are interested in the number of cars that are parked crookedly.

Exercise:

Problem: For every 22 cars, how many would you expect to be parked crookedly, on average?

- A8.25
- B11
- C18
- D7.5

Solution:

A: 8.25

Exercise:

Problem: What is the probability that at least 10 of the 22 cars are parked crookedly.

- A0.1263
- B0.1607
- C0.2870
- D0.8393

Solution:

C: 0.2870

Exercise:**Problem:**

Using a sample of 15 Stanford-Binet IQ scores, we wish to conduct a hypothesis test. Our claim is that the mean IQ score on the Stanford-Binet IQ test is more than 100. It is known that the standard deviation of all Stanford-Binet IQ scores is 15 points. The correct distribution to use for the hypothesis test is:

- A Binomial
- B Student's-t
- C Normal
- D Uniform

Solution:

C: Normal

Questions 11 – 13 refer to the following:

De Anza College keeps statistics on the pass rate of students who enroll in math classes. In a sample of 1795 students enrolled in Math 1A (1st quarter calculus), 1428 passed the course. In a sample of 856 students enrolled in Math 1B (2nd quarter calculus), 662 passed. In general, are the pass rates of Math 1A and Math 1B statistically the same? Let A = the subscript for Math 1A and B = the subscript for Math 1B.

Exercise:

Problem: If you were to conduct an appropriate hypothesis test, the alternate hypothesis would be:

- A $H_a: p_A = p_B$
- B $H_a: p_A > p_B$
- C $H_o: p_A = p_B$
- D $H_a: p_A \neq p_B$

Solution:

D: $H_a: p_A \neq p_B$

Exercise:

Problem: The Type I error is to:

- A conclude that the pass rate for Math 1A is the same as the pass rate for Math 1B when, in fact, the pass rates are different.

- **B**conclude that the pass rate for Math 1A is different than the pass rate for Math 1B when, in fact, the pass rates are the same.
- **C**conclude that the pass rate for Math 1A is greater than the pass rate for Math 1B when, in fact, the pass rate for Math 1A is less than the pass rate for Math 1B.
- **D**conclude that the pass rate for Math 1A is the same as the pass rate for Math 1B when, in fact, they are the same.

Solution:

B: conclude that the pass rate for Math 1A is different than the pass rate for Math 1B when, in fact, the pass rates are the same.

Exercise:

Problem: The correct decision is to:

- **A**reject H_0
- **B**not reject H_0
- **C**There is not enough information given to conduct the hypothesis test

Solution:

B: not reject H_0

Kia, Alejandra, and Iris are runners on the track teams at three different schools. Their running times, in minutes, and the statistics for the track teams at their respective schools, for a one mile run, are given in the table below:

	Running Time	School Average Running Time	School Standard Deviation
Kia	4.9	5.2	.15
Alejandra	4.2	4.6	.25
Iris	4.5	4.9	.12

Exercise:

Problem: Which student is the BEST when compared to the other runners at her school?

- **A**Kia
- **B**Alejandra
- **C**Iris
- **D**Impossible to determine

Solution:

C: Iris

Questions 15 – 16 refer to the following:

The following adult ski sweater prices are from the Gorsuch Ltd. Winter catalog:

{ \$212, \$292, \$278, \$199\$280, \$236 }

Assume the underlying sweater price population is approximately normal. The null hypothesis is that the mean price of adult ski sweaters from Gorsuch Ltd. is at least \$275.

Exercise:

Problem: The correct distribution to use for the hypothesis test is:

- A Normal
- B Binomial
- C Student's-t
- D Exponential

Solution:

C: Student's-t

Exercise:

Problem: The hypothesis test:

- A is two-tailed
- B is left-tailed
- C is right-tailed
- D has no tails

Solution:

B: is left-tailed

Exercise:

Problem:

Sara, a statistics student, wanted to determine the mean number of books that college professors have in their office. She randomly selected 2 buildings on campus and asked each professor in the selected buildings how many books are in his/her office. Sara surveyed 25 professors. The type of sampling selected is a:

- A simple random sampling
- B systematic sampling
- C cluster sampling
- D stratified sampling

Solution:

C: cluster sampling

Exercise:

Problem:

A clothing store would use which measure of the center of data when placing orders for the typical "middle" customer?

- A Mean
- B Median
- C Mode
- D IQR

Solution:

B: Median

Exercise:

Problem: In a hypothesis test, the p-value is

- A the probability that an outcome of the data will happen purely by chance when the null hypothesis is true.
- B called the preconceived alpha.
- C compared to beta to decide whether to reject or not reject the null hypothesis.
- D Answer choices A and B are both true.

Solution:

A: the probability that an outcome of the data will happen purely by chance when the null hypothesis is true.

Questions 20 - 22 refer to the following:

A community college offers classes 6 days a week: Monday through Saturday. Maria conducted a study of the students in her classes to determine how many days per week the students who are in her classes come to campus for classes. In each of her 5 classes she randomly selected 10 students and asked them how many days they come to campus for classes. Each of her classes are the same size. The results of her survey are summarized in the table below.

Number of Days on Campus	Frequency	Relative Frequency	Cumulative Relative Frequency
1	2		
2	12	.24	
3	10	.20	
4			.98
5	0		
6	1	.02	1.00

Exercise:

Problem: Combined with convenience sampling, what other sampling technique did Maria use?

- A simple random
- B systematic
- C cluster
- D stratified

Solution:

D: stratified

Exercise:

Problem: How many students come to campus for classes 4 days a week?

- A 49
- B 25
- C 30
- D 13

Solution:

B: 25

Exercise:

Problem: What is the 60th percentile for the this data?

- A 2
- B 3
- C 4
- D 5

Solution:

C: 4

The next two questions refer to the following:

The following data are the results of a random survey of 110 Reservists called to active duty to increase security at California airports.

Number of Dependents	Frequency
0	11
1	27
2	33
3	20

Number of Dependents	Frequency
4	19

Exercise:

Problem:

Construct a 95% Confidence Interval for the true population mean number of dependents of Reservists called to active duty to increase security at California airports.

- A(1.85, 2.32)
- B(1.80, 2.36)
- C(1.97, 2.46)
- D(1.92, 2.50)

Solution:

A: (1.85, 2.32)

Exercise:

Problem: The 95% confidence Interval above means:

- A5% of Confidence Intervals constructed this way will not contain the true population average number of dependents.
- BWe are 95% confident the true population mean number of dependents falls in the interval.
- CBoth of the above answer choices are correct.
- DNone of the above.

Solution:

C: Both above are correct.

Exercise:

Problem: $X \sim U(4, 10)$. Find the 30th percentile.

- A0.3000
- B3
- C5.8
- D6.1

Solution:

C: 5.8

Exercise:

Problem: If $X \sim \text{Exp}(0.8)$, then $P(x < \mu) =$

- A0.3679
- B0.4727
- C0.6321
- Dcannot be determined

Solution:

C: 0.6321

Exercise:**Problem:**

The lifetime of a computer circuit board is normally distributed with a mean of 2500 hours and a standard deviation of 60 hours. What is the probability that a randomly chosen board will last at most 2560 hours?

- A0.8413
- B0.1587
- C0.3461
- D0.6539

Solution:

A: 0.8413

Exercise:**Problem:**

A survey of 123 Reservists called to active duty as a result of the September 11, 2001, attacks was conducted to determine the proportion that were married. Eighty-six reported being married. Construct a 98% confidence interval for the true population proportion of reservists called to active duty that are married.

- A(0.6030, 0.7954)
- B(0.6181, 0.7802)
- C(0.5927, 0.8057)
- D(0.6312, 0.7672)

Solution:

A: (0.6030, 0.7954)

Exercise:**Problem:**

Winning times in 26 mile marathons run by world class runners average 145 minutes with a standard deviation of 14 minutes. A sample of the last 10 marathon winning times is collected.

Let x = mean winning times for 10 marathons.

The distribution for x is:

- A $N(145, \frac{14}{\sqrt{10}})$
- B $N(145, 14)$
- C t_9
- D t_{10}

Solution:

A: $N = 145, \frac{14}{\sqrt{10}}$

Exercise:

Problem:

Suppose that Phi Beta Kappa honors the top 1% of college and university seniors. Assume that grade point means (G.P.A.) at a certain college are normally distributed with a 2.5 mean and a standard deviation of 0.5. What would be the minimum G.P.A. needed to become a member of Phi Beta Kappa at that college?

- A 3.99
- B 1.34
- C 3.00
- D 3.66

Solution:

D: 3.66

The number of people living on American farms has declined steadily during this century. Here are data on the farm population (in millions of persons) from 1935 to 1980.

Year	1935	1940	1945	1950	1955	1960	1965	1970	1975	1980
Population	32.1	30.5	24.4	23.0	19.1	15.6	12.4	9.7	8.9	7.2

The linear regression equation is $\hat{y} = 1166.93 - 0.5868x$

Exercise:

Problem: What was the expected farm population (in millions of persons) for 1980?

- A 7.2
- B 5.1
- C 6.0
- D 8.0

Solution:

B: 5.1

Exercise:

Problem: In linear regression, which is the best possible SSE?

- A 13.46
- B 18.22
- C 24.05
- D 16.33

Solution:

A: 13.46

Exercise:**Problem:**

In regression analysis, if the correlation coefficient is close to 1 what can be said about the best fit line?

- **A**It is a horizontal line. Therefore, we can not use it.
- **B**There is a strong linear pattern. Therefore, it is most likely a good model to be used.
- **C**The coefficient correlation is close to the limit. Therefore, it is hard to make a decision.
- **D**We do not have the equation. Therefore, we can not say anything about it.

Solution:

B: There is a strong linear pattern. Therefore, it is most likely a good model to be used.

Question 34-36 refer to the following:

A study of the career plans of young women and men sent questionnaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen. Here are the data from the students who responded.

	Female	Male
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

Does the data suggest that there is a relationship between the gender of students and their choice of major?

Exercise:

Problem: The distribution for the test is:

- **A** χ^2_8
- **B** χ^2_3
- **C** t_{721}
- **D** $N(0, 1)$

Solution:

B: χ^2_3

Exercise:

Problem: The expected number of female who choose Finance is :

- A37
- B61
- C60
- D70

Solution:

D: 70

Exercise:

Problem: The p-value is 0.0127 and the level of significance is 0.05. The conclusion to the test is:

- A There is insufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.
- B There is sufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.
- C There is sufficient evidence to conclude that students find Economics very hard.
- D There is in sufficient evidence to conclude that more females prefer Administration than males.

Solution:

B: There is sufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.

Exercise:

Problem:

An agency reported that the work force nationwide is composed of 10% professional, 10% clerical, 30% skilled, 15% service, and 35% semiskilled laborers. A random sample of 100 San Jose residents indicated 15 professional, 15 clerical, 40 skilled, 10 service, and 20 semiskilled laborers. At $\alpha = .10$ does the work force in San Jose appear to be consistent with the agency report for the nation? Which kind of test is it?

- A χ^2 goodness of fit
- B χ^2 test of independence
- C Independent groups proportions
- D Unable to determine

Solution:

A: χ^2 goodness of fit

Practice Final Exam 2

This module is a practice final for an associated elementary statistics textbook, Collaborative Statistics, available for Fall 2008.

Exercise:

Problem:

A study was done to determine the proportion of teenagers that own a car. The population proportion of teenagers that own a car is the

- A statistic
- B parameter
- C population
- D variable

Solution:

B: parameter

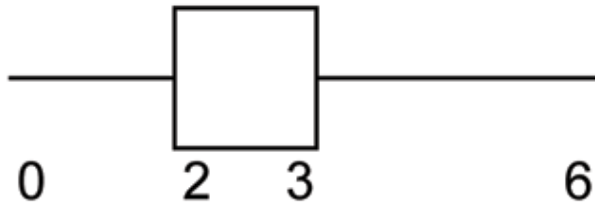
The next two questions refer to the following data:

value	frequency
0	1
1	4
2	7
3	9
6	4

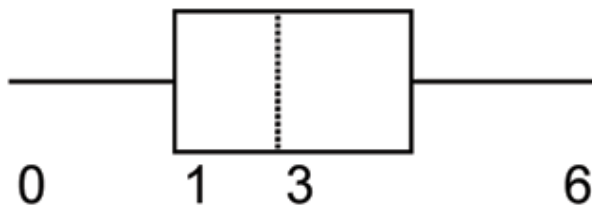
Exercise:

Problem: The box plot for the data is:

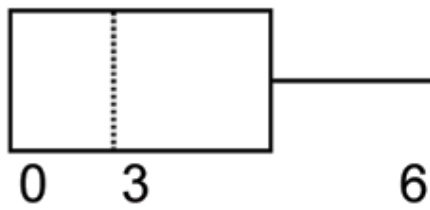
- A



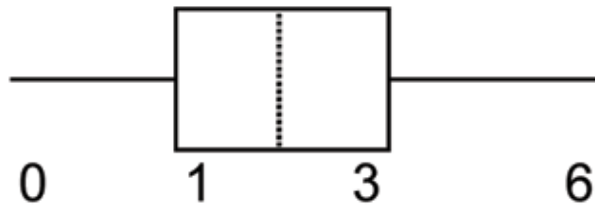
- B



- C



- D



Solution:

A

Exercise:

Problem:

If 6 were added to each value of the data in the table, the 15th percentile of the new list of values is:

- A6
- B1
- C7
- D8

Solution:

C: 7

The next two questions refer to the following situation:

Suppose that the probability of a drought in any independent year is 20%. Out of those years in which a drought occurs, the probability of water rationing is 10%. However, in any year, the probability of water rationing is 5%.

Exercise:

Problem:

What is the probability of both a drought and water rationing occurring?

- A0.05

- B0.01
- C0.02
- D0.30

Solution:

C: 0.02

Exercise:

Problem: Which of the following is true?

- Adrought and water rationing are independent events
- Bdrought and water rationing are mutually exclusive events
- Cnone of the above

Solution:

C: none of the above

The next two questions refer to the following situation:

Suppose that a survey yielded the following data:

gender	apple	pumpkin	pecan
female	40	10	30
male	20	30	10

Favorite Pie Type

Exercise:

Problem:

Suppose that one individual is randomly chosen. The probability that the person's favorite pie is apple or the person is male is:

- A $\frac{40}{60}$
- B $\frac{60}{140}$
- C $\frac{120}{140}$
- D $\frac{100}{140}$

Solution:

D: $\frac{100}{140}$

Exercise:

Problem: Suppose H_0 is: Favorite pie type and gender are independent.

The p-value is:

- A ≈ 0
- B 1
- C 0.05
- D cannot be determined

Solution:

A: ≈ 0

The next two questions refer to the following situation:

Let's say that the probability that an adult watches the news at least once per week is 0.60. We randomly survey 14 people. Of interest is the number that watch the news at least once per week.

Exercise:

Problem: Which of the following statements is FALSE?

- **A** $X \sim B(14, 0.60)$
- **B** The values for x are: $\{ , , , \dots, 14\}$
- **C** $\mu = 8.4$
- **D** $P(X = 5) = 0.0408$

Solution:

B: The values for x are: $\{ , , , \dots, 14\}$

Exercise:

Problem: Find the probability that at least 6 adults watch the news.

- **A** $\frac{6}{14}$
- **B** 0.8499
- **C** 0.9417
- **D** 0.6429

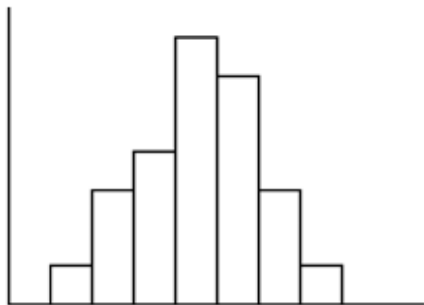
Solution:

C: 0.9417

Exercise:

Problem:

The following histogram is most likely to be a result of sampling from which distribution?



- **A** Chi-Square with $df = 6$
- **B** Exponential

- CUniform
- DBinomial

Solution:

D: Binomial

The ages of campus day and evening students is known to be normally distributed. A sample of 6 campus day and evening students reported their ages (in years) as: {18, 35, 27, 45, 20, 20}

Exercise:

Problem:

What is the error bound for the 90% confidence interval of the true average age?

- A11.2
- B22.3
- C17.5
- D8.7

Solution:

D: 8.7

Exercise:

Problem:

If a normally distributed random variable has $\mu = 0$ and $\sigma = 1$, then 97.5% of the population values lie above:

- A-1.96
- B1.96
- C1
- D-1

Solution:

A: -1.96

The next three questions refer to the following situation:

The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the average amount of money a customer spends in one trip to the supermarket is \$72.

Exercise:

Problem:

What is the probability that one customer spends less than \$72 in one trip to the supermarket?

- A 0.6321
- B 0.5000
- C 0.3714
- D 1

Solution:

A: 0.6321

Exercise:

Problem:

How much money altogether would you expect next 5 customers to spend in one trip to the supermarket (in dollars)?

- A 72
- B $\frac{72^2}{5}$
- C 5184
- D 360

Solution:

D: 360

Exercise:

Problem:

If you want to find the probability that the mean of 50 customers is less than \$60, the distribution to use is:

- A $N(72, 72)$
 - B $N(72, \frac{72}{\sqrt{50}})$
 - C $\text{Exp}(72)$
 - D $\text{Exp}(\frac{1}{72})$
-

Solution:

B: $N(72, \frac{72}{\sqrt{50}})$

The next three questions refer to the following situation:

The amount of time it takes a fourth grader to carry out the trash is uniformly distributed in the interval from 1 to 10 minutes.

Exercise:**Problem:**

What is the probability that a randomly chosen fourth grader takes more than 7 minutes to take out the trash?

- A $\frac{3}{9}$
 - B $\frac{7}{9}$
 - C $\frac{3}{10}$
 - D $\frac{7}{10}$
-

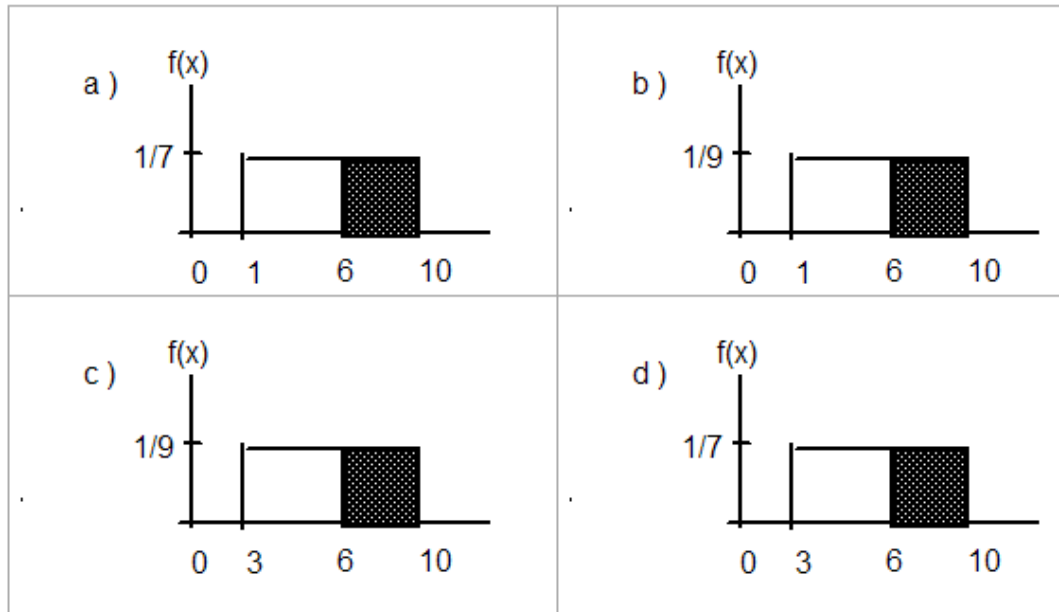
Solution:

A: $\frac{3}{9}$

Exercise:

Problem:

Which graph best shows the probability that a randomly chosen fourth grader takes more than 6 minutes to take out the trash given that he/she has already taken more than 3 minutes?



Solution:

D

Exercise:**Problem:**

We should expect a fourth grader to take how many minutes to take out the trash?

- A 4.5
- B 5.5
- C 5
- D 10

Solution:

B: 5.5

The next three questions refer to the following situation:

At the beginning of the quarter, the amount of time a student waits in line at the campus cafeteria is normally distributed with a mean of 5 minutes and a standard deviation of 1.5 minutes.

Exercise:

Problem: What is the 90th percentile of waiting times (in minutes)?

- A1.28
- B90
- C7.47
- D6.92

Solution:

D: 6.92

Exercise:

Problem: The median waiting time (in minutes) for one student is:

- A5
- B50
- C2.5
- D1.5

Solution:

A: 5

Exercise:

Problem:

Find the probability that the average wait time of 10 students is at most 5.5 minutes.

- A0.6301
- B0.8541
- C0.3694
- D0.1459

Solution:

B: 0.8541

Exercise:

Problem:

A sample of 80 software engineers in Silicon Valley is taken and it is found that 20% of them earn approximately \$50,000 per year. A point estimate for the true proportion of engineers in Silicon Valley who earn \$50,000 per year is:

- A16
- B0.2
- C1
- D0.95

Solution:

B: 0.2

Exercise:

Problem: If $P(Z < z_\alpha) = 0.1587$ where $Z \sim N(0, 1)$, then α is equal to:

- A-1
- B0.1587
- C0.8413
- D1

Solution:

A: -1

Exercise:**Problem:**

A professor tested 35 students to determine their entering skills. At the end of the term, after completing the course, the same test was administered to the same 35 students to study their improvement. This would be a test of:

- A independent groups
- B2 proportions
- C matched pairs, dependent groups
- D exclusive groups

Solution:

C: matched pairs, dependent groups

Exercise:**Problem:**

A math exam was given to all the third grade children attending ABC School. Two random samples of scores were taken.

	n	\bar{x}	s
Boys	55	82	5
Girls	60	86	7

Which of the following correctly describes the results of a hypothesis test of the claim, “There is a difference between the mean scores obtained by third grade girls and boys at the 5 % level of significance”?

- **A** Do not reject H_0 . There is insufficient evidence to conclude that there is a difference in the mean scores.
 - **B** Do not reject H_0 . There is sufficient evidence to conclude that there is a difference in the mean scores.
 - **C** Reject H_0 . There is insufficient evidence to conclude that there is no difference in the mean scores.
 - **D** Reject H_0 . There is sufficient evidence to conclude that there is a difference in the mean scores.
-

Solution:

D: Reject H_0 . There is sufficient evidence to conclude that there is a difference in the mean scores.

Exercise:

Problem:

In a survey of 80 males, 45 had played an organized sport growing up. Of the 70 females surveyed, 25 had played an organized sport growing up. We are interested in whether the proportion for males is higher than the proportion for females. The correct conclusion is:

- **A** There is insufficient information to conclude that the proportion for males is the same as the proportion for females.
 - **B** There is insufficient information to conclude that the proportion for males is not the same as the proportion for females.
 - **C** There is sufficient evidence to conclude that the proportion for males is higher than the proportion for females.
 - **D** Not enough information to determine.
-

Solution:

C: There is sufficient evidence to conclude that the proportion for males is higher than the proportion for females.

Exercise:

Problem:

Note: Chi-Square Test of a Single Variance; Not all classes cover this topic. From past experience, a statistics teacher has found that the average score on a midterm is 81 with a standard deviation of 5.2. This term, a class of 49 students had a standard deviation of 5 on the midterm. Do the data indicate that we should reject the teacher's claim that the standard deviation is 5.2? Use $\alpha = 0.05$.

- A Yes
- B No
- C Not enough information given to solve the problem

Solution:

B: No

Exercise:**Problem:**

Note: F Distribution Test of ANOVA; Not all classes cover this topic. Three loading machines are being compared. Ten samples were taken for each machine. Machine I took an average of 31 minutes to load packages with a standard deviation of 2 minutes. Machine II took an average of 28 minutes to load packages with a standard deviation of 1.5 minutes. Machine III took an average of 29 minutes to load packages with a standard deviation of 1 minute. Find the p-value when testing that the average loading times are the same.

- A the p-value is close to 0
- B p-value is close to 1
- C Not enough information given to solve the problem

Solution:

B: p-value is close to 1.

The next three questions refer to the following situation:

A corporation has offices in different parts of the country. It has gathered the following information concerning the number of bathrooms and the number of employees at seven sites:

Number of employees x	650	730	810	900	102	107	1150
Number of bathrooms y	40	50	54	61	82	110	121

Exercise:

Problem:

Is the correlation between the number of employees and the number of bathrooms significant?

- A Yes
- B No
- C Not enough information to answer question

Solution:

B: No

Exercise:

Problem: The linear regression equation is:

- A $\hat{y} = 0.0094 - 79.96x$

- **B** $\hat{y} = 79.96 + 0.0094x$
- **C** $\hat{y} = 79.96 - 0.0094x$
- **D** $\hat{y} = -0.0094 + 79.96x$

Solution:

C: $\hat{y} = 79.96x - 0.0094$

Exercise:

Problem:

If a site has 1150 employees, approximately how many bathrooms should it have?

- **A** 69
- **B** 91
- **C** 91,954
- **D** We should not be estimating here.

Solution:

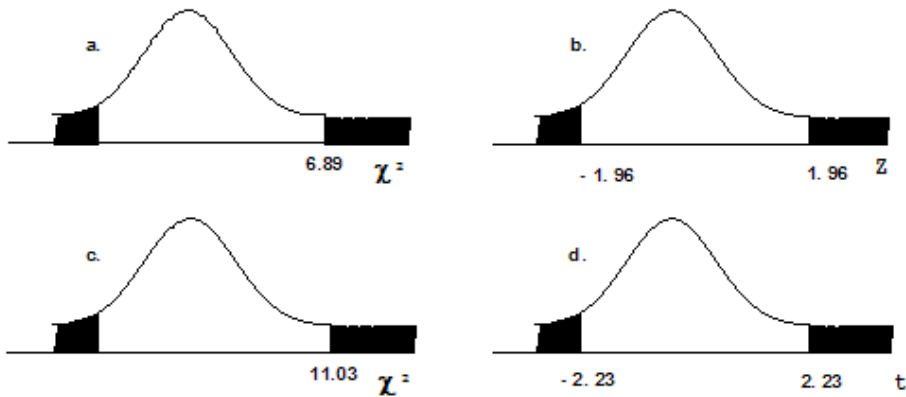
D: We should not be estimating here.

Exercise:

Problem:

Note: Chi-Square Test of a Single Variance; Not all classes cover this topic. Suppose that a sample of size 10 was collected, with $\bar{x} = 4.4$ and $s = 1.4$.

$H_0 : \sigma^2 = 1.6$ vs. $H_a : \sigma^2 \neq 1.6$. Which graph best describes the results of the test?



Solution:

A

Exercise:

Problem:

64 backpackers were asked the number of days their latest backpacking trip was. The number of days is given in the table below:

# of days	1	2	3	4	5	6	7	8
Frequency	5	9	6	12	7	10	5	10

Conduct an appropriate test to determine if the distribution is uniform.

- **A** The p-value is > 0.10 . There is insufficient information to conclude that the distribution is not uniform.
- **B** The p-value is < 0.01 . There is sufficient information to conclude the distribution is not uniform.
- **C** The p-value is between 0.01 and 0.10, but without alpha (α) there is not enough information
- **D** There is no such test that can be conducted.

Solution:

A: The p-value is > 0.10 . There is insufficient information to conclude that the distribution is not uniform.

Exercise:**Problem:**

Note: F Distribution test of One-Way ANOVA; Not all classes cover this topic. Which of the following statements is true when using one-way ANOVA?

- A The populations from which the samples are selected have different distributions.
- B The sample sizes are large.
- C The test is to determine if the different groups have the same means.
- D There is a correlation between the factors of the experiment.

Solution:

C: The test is to determine if the different groups have the same means.

Data Sets

This module provides data sets for use with the Collaborative Statistics textbook/collection. Data sets include a series of recorded motorcycle race and practice lap times as well as IPO stock prices.

Lap Times

The following tables provide lap times from Terri Vogel's Log Book. Times are recorded in seconds for 2.5-mile laps completed in a series of races and practice runs.

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Race 1	135	130	131	132	130	131	133
Race 2	134	131	131	129	128	128	129
Race 3	129	128	127	127	130	127	129
Race 4	125	125	126	125	124	125	125
Race 5	133	132	132	132	131	130	132
Race 6	130	130	130	129	129	130	129
Race 7	132	131	133	131	134	134	131

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Race 8	127	128	127	130	128	126	128
Race 9	132	130	127	128	126	127	124
Race 10	135	131	131	132	130	131	130
Race 11	132	131	132	131	130	129	129
Race 12	134	130	130	130	131	130	130
Race 13	128	127	128	128	128	129	128
Race 14	132	131	131	131	132	130	130
Race 15	136	129	129	129	129	129	129
Race 16	129	129	129	128	128	129	129
Race 17	134	131	132	131	132	132	132
Race 18	129	129	130	130	133	133	127
Race 19	130	129	129	129	129	129	128

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Race 20	131	128	130	128	129	130	130

Race Lap Times (in Seconds)

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Practice 1	142	143	180	137	134	134	172
Practice 2	140	135	134	133	128	128	131
Practice 3	130	133	130	128	135	133	133
Practice 4	141	136	137	136	136	136	145
Practice 5	140	138	136	137	135	134	134
Practice 6	142	142	139	138	129	129	127
Practice 7	139	137	135	135	137	134	135
Practice 8	143	136	134	133	134	133	132

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Practice 9	135	134	133	133	132	132	133
Practice 10	131	130	128	129	127	128	127
Practice 11	143	139	139	138	138	137	138
Practice 12	132	133	131	129	128	127	126
Practice 13	149	144	144	139	138	138	137
Practice 14	133	132	137	133	134	130	131
Practice 15	138	136	133	133	132	131	131

Practice Lap Times (in Seconds)

Stock Prices

The following table lists initial public offering (IPO) stock prices for all 1999 stocks that at least doubled in value during the first day of trading. This is historical data.

\$17.00	\$23.00	\$14.00	\$16.00	\$12.00	\$26.00

\$20.00	\$22.00	\$14.00	\$15.00	\$22.00	\$18.00
\$18.00	\$21.00	\$21.00	\$19.00	\$15.00	\$21.00
\$18.00	\$17.00	\$15.00	\$25.00	\$14.00	\$30.00
\$16.00	\$10.00	\$20.00	\$12.00	\$16.00	\$17.44
\$16.00	\$14.00	\$15.00	\$20.00	\$20.00	\$16.00
\$17.00	\$16.00	\$15.00	\$15.00	\$19.00	\$48.00
\$16.00	\$18.00	\$9.00	\$18.00	\$18.00	\$20.00
\$8.00	\$20.00	\$17.00	\$14.00	\$11.00	\$16.00
\$19.00	\$15.00	\$21.00	\$12.00	\$8.00	\$16.00
\$13.00	\$14.00	\$15.00	\$14.00	\$13.41	\$28.00
\$21.00	\$17.00	\$28.00	\$17.00	\$19.00	\$16.00
\$17.00	\$19.00	\$18.00	\$17.00	\$15.00	
\$14.00	\$21.00	\$12.00	\$18.00	\$24.00	
\$15.00	\$23.00	\$14.00	\$16.00	\$12.00	
\$24.00	\$20.00	\$14.00	\$14.00	\$15.00	
\$14.00	\$19.00	\$16.00	\$38.00	\$20.00	
\$24.00	\$16.00	\$8.00	\$18.00	\$17.00	
\$16.00	\$15.00	\$7.00	\$19.00	\$12.00	
\$8.00	\$23.00	\$12.00	\$18.00	\$20.00	
\$21.00	\$34.00	\$16.00	\$26.00	\$14.00	

IPO Offer Prices

Note: *Data compiled by Jay R. Ritter of Univ. of Florida using data from Securities Data Co. and Bloomberg.*

Group Project: Univariate Data

In this project, students will design and carry out a survey, analyze the data, and graphically display the results.

Student Learning Objectives

- The student will design and carry out a survey.
- The student will analyze and graphically display the results of the survey.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

- ____Decide what data you are going to study.

Note:Here are two examples, but you may **NOT** use them: number of M&M's per small bag, number of pencils students have in their backpacks.

- ____Are your data discrete or continuous? How do you know?
- ____Decide how you are going to collect the data (for instance, buy 30 bags of M&M's; collect data from the World Wide Web).
- ____Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. What method did you use? Why did you pick that method?
- ____Conduct your survey. **Your data size must be at least 30.**
- ____Summarize your data in a chart with columns showing **data value, frequency, relative frequency and cumulative relative frequency.**
- ____Answer the following (rounded to 2 decimal places):

- $1x =$
 - $2s =$
 - 3First quartile =
 - 4Median =
 - 570th percentile =
- ____ What value is 2 standard deviations above the mean?
 - ____ What value is 1.5 standard deviations below the mean?
 - ____ Construct a histogram displaying your data.
 - ____ In complete sentences, describe the shape of your graph.
 - ____ Do you notice any potential outliers? If so, what values are they?
Show your work in how you used the potential outlier formula in Chapter 2 (since you have univariate data) to determine whether or not the values might be outliers.
 - ____ Construct a box plot displaying your data.
 - ____ Does the middle 50% of the data appear to be concentrated together or spread apart? Explain how you determined this.
 - ____ Looking at both the histogram and the box plot, discuss the distribution of your data.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

- ____ **Cover sheet:** name, class time, and name of your study
- ____ **Summary page:** This should contain paragraphs written with complete sentences. It should include answers to all the questions above. It should also include statements describing the population under study, the sample, a parameter or parameters being studied, and the statistic or statistics produced.
- ____ **URL** for data, if your data are from the World Wide Web.
- ____ **Chart of data, frequency, relative frequency and cumulative relative frequency.**
- ____ **Page(s) of graphs:** histogram and box plot.

Group Project: Continuous Distributions and Central Limit Theorem

In this project, students will identify and analyze a continuous data set, determine which distribution model most closely describes the data, and calculate probabilities.

Student Learning Objectives

- The student will collect a sample of continuous data.
- The student will attempt to fit the data sample to various distribution models.
- The student will validate the Central Limit Theorem.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

Part I: Sampling

- ____Decide what **continuous** data you are going to study. (Here are two examples, but you may NOT use them: the amount of money a student spends on college supplies this term or the length of a long distance telephone call.)
- ____Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. What method did you use? Why did you pick that method?
- ____Conduct your survey. Gather **at least 150 pieces of continuous quantitative data**.
- ____Define (in words) the random variable for your data. $X =$

- ____Create 2 lists of your data: (1) unordered data, (2) in order of smallest to largest.
- ____Find the sample mean and the sample standard deviation (rounded to 2 decimal places).

- $1\bar{x} =$
- $2s =$
- ____ Construct a histogram of your data containing 5 - 10 intervals of equal width. The histogram should be a representative display of your data. Label and scale it.

Part II: Possible Distributions

- ____ Suppose that X followed the theoretical distributions below. Set up each distribution using the appropriate information from your data.
- ____ Uniform: $X \sim U$ ____ Use the lowest and highest values as a and b .
- ____ Exponential: $X \sim \text{Exp}$ ____ Use x to estimate μ .
- ____ Normal: $X \sim N$ ____ Use x to estimate for μ and s to estimate for σ .
- ____ **Must** your data fit one of the above distributions? Explain why or why not.
- ____ **Could** the data fit 2 or 3 of the above distributions (at the same time)? Explain.
- ____ Calculate the value k (an X value) that is 1.75 standard deviations above the sample mean. $k =$ ____ (rounded to 2 decimal places)
Note: $k = \bar{x} + (1.75)*s$
- ____ Determine the relative frequencies (RF) rounded to 4 decimal places.
 - $1\text{RF} = \frac{\text{frequency}}{\text{total number surveyed}}$
 - $2\text{RF}(X < k) =$
 - $3\text{RF}(X > k) =$
 - $4\text{RF}(X = k) =$

Use a separate piece of paper for EACH distribution (uniform, exponential, normal) to respond to the following questions.

Note: You should have one page for the uniform, one page for the exponential, and one page for the normal

- ____ State the distribution: $X \sim$ ____
- ____ Draw a graph for each of the three theoretical distributions. Label the axes and mark them appropriately.
- ____ Find the following theoretical probabilities (rounded to 4 decimal places).
 - 1 $P(X < k) =$
 - 2 $P(X > k) =$
 - 3 $P(X = k) =$
- ____ Compare the relative frequencies to the corresponding probabilities. Are the values close?
- ____ Does it appear that the data fit the distribution well? Justify your answer by comparing the probabilities to the relative frequencies, and the histograms to the theoretical graphs.

Part III: CLT Experiments

- ____ From your original data (before ordering), use a random number generator to pick 40 samples of size 5. For each sample, calculate the average.
- ____ On a separate page, attached to the summary, include the 40 samples of size 5, along with the 40 sample averages.
- ____ List the 40 averages in order from smallest to largest.
- ____ Define the random variable, X , in words. $X =$
- ____ State the approximate theoretical distribution of X . $X \sim$
- ____ Base this on the mean and standard deviation from your original data.
- ____ Construct a histogram displaying your data. Use 5 to 6 intervals of equal width. Label and scale it.
- Calculate the value k (an X value) that is 1.75 standard deviations above the sample mean. $k =$ ____ (rounded to 2 decimal places)

- Determine the relative frequencies (RF) rounded to 4 decimal places.
 - $1RF(X < k) =$
 - $2RF(X > k) =$
 - $3RF(X = k) =$
- Find the following theoretical probabilities (rounded to 4 decimal places).
 - $1P(X < k) =$
 - $2P(X > k) =$
 - $3P(X = k) =$
- _____ Draw the graph of the theoretical distribution of X .
- _____ Answer the questions below.
- _____ Compare the relative frequencies to the probabilities. Are the values close?
- _____ Does it appear that the data of averages fit the distribution of \bar{X} well? Justify your answer by comparing the probabilities to the relative frequencies, and the histogram to the theoretical graph.
- _____ In 3 - 5 complete sentences for each, answer the following questions. Give thoughtful explanations.
- _____ In summary, do your original data seem to fit the uniform, exponential, or normal distributions? Answer why or why not for each distribution. If the data do not fit any of those distributions, explain why.
- _____ What happened to the shape and distribution when you averaged your data? **In theory**, what should have happened? In theory, would “it” always happen? Why or why not?
- _____ Were the relative frequencies compared to the theoretical probabilities closer when comparing the X or \bar{X} distributions? Explain your answer.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

- ____ **Cover sheet:** name, class time, and name of your study
- ____ **Summary pages:** These should contain several paragraphs written with complete sentences that describe the experiment, including what you studied and your sampling technique, as well as answers to all of the questions above.
- ____ **URL** for data, if your data are from the World Wide Web.
- ____ **Pages, one for each theoretical distribution,** with the distribution stated, the graph, and the probability questions answered
- ____ **Pages of the data requested**
- ____ **All graphs required**

Partner Project: Hypothesis Testing - Article

In this project, students will identify real-world examples of hypothesis testing in the media. Students will then conduct their own survey and compare results.

Student Learning Objectives

- The student will identify a hypothesis testing problem in print.
- The student will conduct a survey to verify or dispute the results of the hypothesis test.
- The student will summarize the article, analysis, and conclusions in a report.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

- ____ **Find an article** in a newspaper, magazine or on the internet which makes a claim about **ONE** population mean or **ONE** population proportion. The claim may be based upon a survey that the article was reporting on. Decide whether this claim is the null or alternate hypothesis.
- ____ **Copy or print out the article** and include a copy in your project, along with the source.
- ____ **State how you will collect your data.** (Convenience sampling is not acceptable.)
- ____ **Conduct your survey. You must have more than 50 responses in your sample.** When you hand in your final project, attach the tally sheet or the packet of questionnaires that you used to collect data. Your data must be real.
- ____ **State the statistics** that are a result of your data collection: sample size, sample mean, and sample standard deviation, OR sample size and number of successes.
- ____ **Make 2 copies of the appropriate solution sheet.**
- ____ **Record the hypothesis test** on the solution sheet, based on your experiment. **Do a DRAFT solution** first on one of the solution sheets

and check it over carefully. Have a classmate check your solution to see if it is done correctly. Make your decision using a 5% level of significance. Include the 95% confidence interval on the solution sheet.

- ____ **Create a graph that illustrates your data.** This may be a pie or bar chart or may be a histogram or box plot, depending on the nature of your data. Produce a graph that makes sense for your data and gives useful visual information about your data. You may need to look at several types of graphs before you decide which is the most appropriate for the type of data in your project.
- ____ **Write your summary** (in complete sentences and paragraphs, with proper grammar and correct spelling) that describes the project. The summary **MUST** include:
 - **1**Brief discussion of the article, including the source.
 - **2**Statement of the claim made in the article (one of the hypotheses).
 - **3**Detailed description of how, where, and when you collected the data, including the sampling technique. Did you use cluster, stratified, systematic, or simple random sampling (using a random number generator)? As stated above, convenience sampling is not acceptable.
 - **4**Conclusion about the article claim in light of your hypothesis test. This is the conclusion of your hypothesis test, stated in words, in the context of the situation in your project in sentence form, as if you were writing this conclusion for a non-statistician.
 - **5**Sentence interpreting your confidence interval in the context of the situation in your project.

Assignment Checklist

Turn in the following typed (12 point) and stapled packet for your final project:

- ____ **Cover sheet** containing your name(s), class time, and the name of your study.
- ____ **Summary**, which includes all items listed on summary checklist.

- ____ **Solution sheet** neatly and completely filled out. The solution sheet does not need to be typed.
- ____ **Graphic representation of your data**, created following the guidelines discussed above. Include only graphs which are appropriate and useful.
- ____ **Raw data collected AND a table summarizing the sample data** (n , \bar{x} and s ; or x , n , and p' , as appropriate for your hypotheses). The raw data does not need to be typed, but the summary does. Hand in the data as you collected it. (Either attach your tally sheet or an envelope containing your questionnaires.)

Partner Project: Hypothesis Testing - Word Problem

In this project, students will write, edit, and solve a hypothesis testing word problem.

Student Learning Objectives

- The student will write, edit, and solve a hypothesis testing word problem.

Instructions

Write an original hypothesis testing problem for either **ONE** population mean or **ONE** population proportion. As you complete each task, check it off. Answer all questions in your summary. Look at the homework for the Hypothesis Testing: Single Mean and Single Proportion chapter for examples (poems, two acts of a play, a work related problem). The problems with names attached to them are problems written by students in past quarters. Some other examples that are not in the homework include: a soccer hypothesis testing poster, a cartoon, a news reports, a children's story, a song.

- ____ Your problem must be original and creative. It also must be in proper English. If English is difficult for you, have someone edit your problem.
- ____ Your problem must be at least $\frac{1}{2}$ page, typed and singled spaced. This **DOES NOT** include the data. Data will make the problem longer and that is fine. For this problem, the data and story may be real or fictional.
- ____ In the narrative of the problem, make it very clear what the null and alternative hypotheses are.
- ____ Your sample size must be **LARGER THAN 50** (even if it is fictional).
- ____ State in your problem how you will collect your data.
- ____ Include your data with your word problem.
- ____ State the statistics that are a result of your data collection: sample size, sample mean, and sample standard deviation, OR sample size and number of successes.

- ____ Create a graph that illustrates your problem. This may be a pie or bar chart or may be a histogram or box plot, depending on the nature of your data. Produce a graph that makes sense for your data and gives useful visual information about your data. You may need to look at several types of graphs before you decide which is the most appropriate for your problem.
- ____ Make 2 copies of the appropriate solution sheet.
- ____ Record the hypothesis test on the solution sheet, based on your problem. Do a **DRAFT** solution first on one of the solution sheets and check it over carefully. Make your decision using a 5% level of significance. Include the 95% confidence interval on the solution

Assignment Checklist

You need to turn in the following typed (12 point) and stapled packet for your final project:

- ____ **Cover sheet** containing your name, the name of your problem, and the date
- ____ **The problem**
- ____ **Data for the problem**
- ____ **Solution sheet** neatly and completely filled out. The solution sheet does not need to be typed.
- ____ **Graphic representation of the data**, created following the guidelines discussed above. Include only graphs that are appropriate and useful.
- ____ **Sentences interpreting the results of the hypothesis test and the confidence interval** in the context of the situation in the project.

Group Project: Bivariate Data, Linear Regression, and Univariate Data

In this project, students will collect a sample of bivariate data and analyze the information. Students will be asked to describe the center and spread of the data, determine the goodness of fit of a linear regression model, and analyze the relationship between the variables.

Student Learning Objectives

- The students will collect a bivariate data sample through the use of appropriate sampling techniques.
- The student will attempt to fit the data to a linear model.
- The student will determine the appropriateness of linear fit of the model.
- The student will analyze and graph univariate data.

Instructions

1. As you complete each task below, check it off. Answer all questions in your introduction or summary.
2. Check your course calendar for intermediate and final due dates.
3. Graphs may be constructed by hand or by computer, unless your instructor informs you otherwise. All graphs must be neat and accurate.
4. All other responses must be done on the computer.
5. Neatness and quality of explanations are used to determine your final grade.

Part I: Bivariate Data

Introduction

- ____ State the bivariate data your group is going to study.

Note: Here are two examples, but you may **NOT** use them: height vs. weight and age vs. running distance.

- ____ Describe how your group is going to collect the data (for instance, collect data from the web, survey students on campus).
- ____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random sampling (using a random number generator) sampling. Convenience sampling is **NOT** acceptable.
- ____ Conduct your survey. Your number of pairs must be at least 30.
- ____ Print out a copy of your data.

Analysis

- ____ On a separate sheet of paper construct a scatter plot of the data. Label and scale both axes.
- ____ State the least squares line and the correlation coefficient.
- ____ On your scatter plot, in a different color, construct the least squares line.
- ____ Is the correlation coefficient significant? Explain and show how you determined this.
- ____ Interpret the slope of the linear regression line in the context of the data in your project. Relate the explanation to your data, and quantify what the slope tells you.
- ____ Does the regression line seem to fit the data? Why or why not? If the data does not seem to be linear, explain if any other model seems to fit the data better.
- ____ Are there any outliers? If so, what are they? Show your work in how you used the potential outlier formula in the Linear Regression and Correlation chapter (since you have bivariate data) to determine whether or not any pairs might be outliers.

Part II: Univariate Data

In this section, you will use the data for **ONE** variable only. Pick the variable that is more interesting to analyze. For example: if your

independent variable is sequential data such as year with 30 years and one piece of data per year, your x-values might be 1971, 1972, 1973, 1974, ..., 2000. This would not be interesting to analyze. In that case, choose to use the dependent variable to analyze for this part of the project.

- ____ Summarize your data in a chart with columns showing data value, frequency, relative frequency, and cumulative relative frequency.
- ____ Answer the following, rounded to 2 decimal places:
 - **1** Sample mean =
 - **2** Sample standard deviation =
 - **3** First quartile =
 - **4** Third quartile =
 - **5** Median =
 - **6** 70th percentile =
 - **7** Value that is 2 standard deviations above the mean =
 - **8** Value that is 1.5 standard deviations below the mean =
- ____ Construct a histogram displaying your data. Group your data into 6 – 10 intervals of equal width. Pick regularly spaced intervals that make sense in relation to your data. For example, do NOT group data by age as 20-26, 27-33, 34-40, 41-47, 48-54, 55-61 . . . Instead, maybe use age groups 19.5-24.5, 24.5-29.5, . . . or 19.5-29.5, 29.5-39.5, 39.5-49.5, . . .
- ____ In complete sentences, describe the shape of your histogram.
- ____ Are there any potential outliers? Which values are they? Show your work and calculations as to how you used the potential outlier formula in chapter 2 (since you are now using univariate data) to determine which values might be outliers.
- ____ Construct a box plot of your data.
- ____ Does the middle 50% of your data appear to be concentrated together or spread out? Explain how you determined this.
- ____ Looking at both the histogram AND the box plot, discuss the distribution of your data. For example: how does the spread of the middle 50% of your data compare to the spread of the rest of the data represented in the box plot; how does this correspond to your

description of the shape of the histogram; how does the graphical display show any outliers you may have found; does the histogram show any gaps in the data that are not visible in the box plot; are there any interesting features of your data that you should point out.

Due Dates

- Part I, Intro: _____ (keep a copy for your records)
- Part I, Analysis: _____ (keep a copy for your records)
- Entire Project, typed and stapled: _____
 - _____ Cover sheet: names, class time, and name of your study.
 - _____ Part I: label the sections “Intro” and “Analysis.”
 - _____ Part II:
 - _____ Summary page containing several paragraphs written in complete sentences describing the experiment, including what you studied and how you collected your data. The summary page should also include answers to ALL the questions asked above.
 - _____ All graphs requested in the project.
 - _____ All calculations requested to support questions in data.
 - _____ Description: what you learned by doing this project, what challenges you had, how you overcame the challenges.

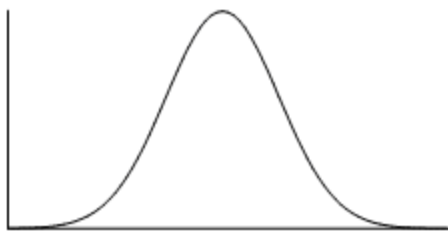
Note: Include answers to ALL questions asked, even if not explicitly repeated in the items above.

Solution Sheet: Hypothesis Testing for Single Mean and Single Proportion
This module provides a solution sheet for the Hypothesis Testing: Single Mean and Single Proportion chapter of the Collaborative Statistics textbook/collection.

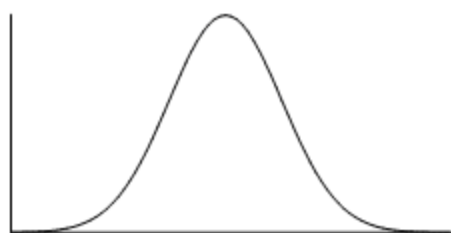
Class Time:

Name:

- **a** H_o :
- **b** H_a :
- **c** In words, **CLEARLY** state what your random variable X or P' represents.
- **d** State the distribution to use for the test.
- **e** What is the test statistic?
- **f** What is the p -value? In 1 – 2 complete sentences, explain what the p -value means for this problem.
- **g** Use the previous information to sketch a picture of this situation. **CLEARLY**, label and scale the horizontal axis and shade the region(s) corresponding to the p -value.



- **h** Indicate the correct decision (“reject” or “do not reject” the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.
 - **i** Alpha:
 - **ii** Decision:
 - **iii** Reason for decision:
 - **iv** Conclusion:
- **i** Construct a 95% Confidence Interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the Confidence Interval.



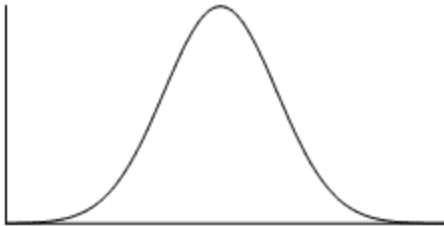
Solution Sheet: Hypothesis Testing for Two Means, Paired Data, and Two Proportions

This module provides a solution sheet for the Hypothesis Testing: Two Means, Paired Data, Two Proportions chapter of the Collaborative Statistics textbook/collection.

Class Time:

Name:

- **a** H_o : _____
- **b** H_a : _____
- **c** In words, **clearly** state what your random variable $X_1 - X_2$, $P_1' - P_2'$ or X_d represents.
- **d** State the distribution to use for the test.
- **e** What is the test statistic?
- **f** What is the p -value? In 1 – 2 complete sentences, explain what the p -value means for this problem.
- **g** Use the previous information to sketch a picture of this situation. **CLEARLY** label and scale the horizontal axis and shade the region(s) corresponding to the p -value.



- **h** Indicate the correct decision (“reject” or “do not reject” the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.
 - **i** Alpha:
 - **ii** Decision:
 - **iii** Reason for decision:
 - **iv** Conclusion:
- **i** In complete sentences, explain how you determined which distribution to use.

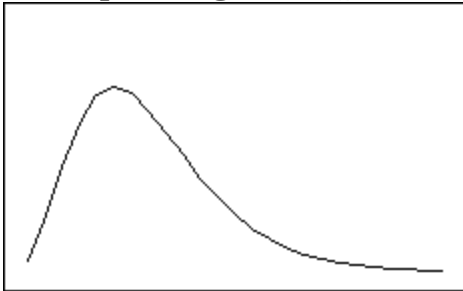
Solution Sheet: The Chi-Square Distribution

This module provides a solution sheet for the Chi-Square Distribution chapter of the Collaborative Statistics textbook/collection.

Class Time:

Name:

- **a** H_o : _____
- **b** H_a : _____
- **c** What are the degrees of freedom?
- **d** State the distribution to use for the test.
- **e** What is the test statistic?
- **f** What is the p -value? In 1 – 2 complete sentences, explain what the p -value means for this problem.
- **g** Use the previous information to sketch a picture of this situation. **Clearly** label and scale the horizontal axis and shade the region(s) corresponding to the p -value.



- **h** Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using **complete sentences**.
 - **i** Alpha:
 - **ii** Decision:
 - **iii** Reason for decision:
 - **iv** Conclusion:

Solution Sheet: F Distribution and ANOVA

This module provides a solution sheet for use with the F Distribution and One-Way ANOVA chapter of the Collaborative Statistics textbook/collection.

Class Time:

Name:

- **a** H_o :
- **b** H_a :
- **c** $\text{cdf}(n) = \underline{\hspace{2cm}}$ $\text{df}(d) = \underline{\hspace{2cm}}$
- **d** State the distribution to use for the test.
- **e** What is the test statistic?
- **f** What is the p -value?
- **g** Use the previous information to sketch a picture of this situation. **Clearly** label and scale the horizontal axis and shade the region(s) corresponding to the p -value.



- **h** Indicate the correct decision (“reject” or “do not reject” the null hypothesis) and write appropriate conclusions, using **complete sentences**.
 - **i** Alpha:
 - **ii** Decision:
 - **iii** Reason for decision:
 - **iv** Conclusion:

English Phrases Written Mathematically

This module provides an overview of commonly used phrases in statistics and their mathematical equivalents.

English Phrases Written Mathematically

When the English says:	Interpret this as:
X is at least 4.	$X \geq 4$
The minimum of X is 4.	$X \geq 4$
X is no less than 4.	$X \geq 4$
X is greater than or equal to 4.	$X \geq 4$
X is at most 4.	$X \leq 4$
The maximum of X is 4.	$X \leq 4$
X is no more than 4.	$X \leq 4$
X is less than or equal to 4.	$X \leq 4$
X does not exceed 4.	$X \leq 4$
X is greater than 4.	$X > 4$

When the English says:	Interpret this as:
X is more than 4.	$X > 4$
X exceeds 4.	$X > 4$
X is less than 4.	$X < 4$
There are fewer X than 4.	$X < 4$
X is 4.	$X = 4$
X is equal to 4.	$X = 4$
X is the same as 4.	$X = 4$
X is not 4.	$X \neq 4$
X is not equal to 4.	$X \neq 4$
X is not the same as 4.	$X \neq 4$
X is different than 4.	$X \neq 4$

Symbols and their Meanings

This module defines symbols used throughout the Collaborative Statistics textbook.

Chapter (1st used)	Symbol	Spoken	Meaning
Sampling and Data	$\sqrt{\quad}$	The square root of	same
Sampling and Data		Pi	3.14159... (a specific number)
Descriptive Statistics	Q1	Quartile one	the first quartile
Descriptive Statistics	Q2	Quartile two	the second quartile
Descriptive Statistics	Q3	Quartile three	the third quartile
Descriptive Statistics	IQR	inter-quartile range	Q3- Q1=IQR
Descriptive Statistics		x-bar	sample mean
Descriptive Statistics		mu	population mean

Chapter (1st used)	Symbol	Spoken	Meaning
Descriptive Statistics	s_x	s	sample standard deviation
Descriptive Statistics	s^2	s-squared	sample variance
Descriptive Statistics	σ_x	sigma	population standard deviation
Descriptive Statistics	σ^2	sigma-squared	population variance
Descriptive Statistics		capital sigma	sum
Probability Topics	$\{ \}$	brackets	set notation
Probability Topics		S	sample space
Probability Topics		Event A	event A
Probability Topics	$()$	probability of A	probability of A occurring

Chapter (1st used)	Symbol	Spoken	Meaning
Probability Topics	$(\quad \quad)$	probability of A given B	prob. of A occurring given B has occurred
Probability Topics	$(\quad \text{or} \quad)$	prob. of A or B	prob. of A or B or both occurring
Probability Topics	$(\quad \text{and} \quad)$	prob. of A and B	prob. of both A and B occurring (same time)
Probability Topics	A'	A-prime, complement of A	complement of A, not A
Probability Topics	(A')	prob. of complement of A	same
Probability Topics	1	green on first pick	same
Probability Topics	$(\quad 1)$	prob. of green on first pick	same
Discrete Random Variables	PDF	prob. distribution function	same

Chapter (1st used)	Symbol	Spoken	Meaning
Discrete Random Variables		X	the random variable X
Discrete Random Variables	$X \sim$	the distribution of X	same
Discrete Random Variables		binomial distribution	same
Discrete Random Variables		geometric distribution	same
Discrete Random Variables		hypergeometric dist.	same
Discrete Random Variables		Poisson dist.	same
Discrete Random Variables		Lambda	average of Poisson distribution
Discrete Random Variables	\geq	greater than or equal to	same

Chapter (1st used)	Symbol	Spoken	Meaning
Discrete Random Variables	\leq	less than or equal to	same
Discrete Random Variables	$=$	equal to	same
Discrete Random Variables	\neq	not equal to	same
Continuous Random Variables	$()$	f of x	function of x
Continuous Random Variables	pdf	prob. density function	same
Continuous Random Variables		uniform distribution	same
Continuous Random Variables	Exp	exponential distribution	same
Continuous Random Variables		k	critical value

Chapter (1st used)	Symbol	Spoken	Meaning
Continuous Random Variables	$() =$	f of x equals	same
Continuous Random Variables		m	decay rate (for exp. dist.)
The Normal Distribution		normal distribution	same
The Normal Distribution		z-score	same
The Normal Distribution		standard normal dist.	same
The Central Limit Theorem	CLT	Central Limit Theorem	same
The Central Limit Theorem		X-bar	the random variable X- bar
The Central Limit Theorem		mean of X	the average of X
The Central Limit Theorem		mean of X-bar	the average of X-bar

Chapter (1st used)	Symbol	Spoken	Meaning
The Central Limit Theorem		standard deviation of X	same
The Central Limit Theorem		standard deviation of X-bar	same
The Central Limit Theorem		sum of X	same
The Central Limit Theorem		sum of x	same
Confidence Intervals	CL	confidence level	same
Confidence Intervals	CI	confidence interval	same
Confidence Intervals	EBM	error bound for a mean	same
Confidence Intervals	EBP	error bound for a proportion	same
Confidence Intervals		student-t distribution	same
Confidence Intervals	df	degrees of freedom	same

Chapter (1st used)	Symbol	Spoken	Meaning
Confidence Intervals	\bar{z}	student-t with a/2 area in right tail	same
Confidence Intervals	p'	p-prime; p-hat	sample proportion of success
Confidence Intervals	$q' \wedge$	q-prime; q-hat	sample proportion of failure
Hypothesis Testing	0	H-naught, H- sub 0	null hypothesis
Hypothesis Testing		H-a, H-sub a	alternate hypothesis
Hypothesis Testing	1	H-1, H-sub 1	alternate hypothesis
Hypothesis Testing		alpha	probability of Type I error
Hypothesis Testing		beta	probability of Type II error
Hypothesis Testing	$-$	X1-bar minus X2-bar	difference in sample means

Chapter (1st used)	Symbol	Spoken	Meaning
	$\mu_1 - \mu_2$	mu-1 minus mu-2	difference in population means
	$p_1 - p_2$	P1-prime minus P2-prime	difference in sample proportions
	$p_1 - p_2$	p1 minus p2	difference in population proportions
Chi-Square Distribution	χ^2	Ky-square	Chi-square
		Observed	Observed frequency
		Expected	Expected frequency
Linear Regression and Correlation	$y = a + bx$	y equals a plus b-x	equation of a line
	\hat{y}	y-hat	estimated value of y
		correlation coefficient	same

Chapter (1st used)	Symbol	Spoken	Meaning
		error	same
	SSE	Sum of Squared Errors	same
	1.9	1.9 times s	cut-off value for outliers
F- Distribution and ANOVA		F-ratio	F ratio

Symbols and their Meanings

Formulas

This module provides an overview of Statistics Formulas used as a part of Collaborative Statistics collection (col10522) by Barbara Illowsky and Susan Dean.

Formula

Factorial

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 1$$

Formula

Combinations

$${}^n C_r = \frac{n!}{r!(n-r)!}$$

Formula

Binomial Distribution

$$X \sim B(n, p)$$

$$P(X = x) = {}^n C_x p^x q^{n-x}, \text{ for } x = 0, 1, 2, \dots, n$$

Formula

Geometric Distribution

$$X \sim G(p)$$

$$P(X = x) = q^x p, \text{ for } x = 1, 2, 3, \dots$$

Formula

Hypergeometric Distribution

$$r \leq b \leq n$$

$$P(X = x) = \frac{{}^r C_x {}^{n-r} C_{n-x}}{{}^n C_n}$$

Formula

Poisson Distribution

$$\mu$$

$$P(X \leq x) = \frac{\mu^x e^{-\mu}}{x}$$

Formula

Uniform Distribution

$$X \sim U(a, b)$$

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b$$

Formula

Exponential Distribution

$$X \sim \text{Exp}(m)$$

$$f(x) = me^{-mx}, \quad m > 0, x \geq 0$$

Formula

Normal Distribution

$$X \sim N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sigma \sqrt{\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

Formula

Gamma Function

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx, \quad \text{Re}(z) > 0$$

$$\Gamma(n) = (n-1)!$$

$$\Gamma(m) = (m-1)! \quad \text{for } m, \text{ a nonnegative integer}$$

$$\text{otherwise: } \Gamma(a) \Gamma(b) = \Gamma(a+b) \int_0^1 t^{a-1} (1-t)^{b-1} dt$$

Formula

Student-t Distribution

$$X \sim t_\nu$$

$$f(x) = \frac{\frac{x}{n} \frac{n}{\Gamma(\frac{n}{2})}}{\pi \Gamma(\frac{n}{2})}$$

$$X = \frac{Z}{\frac{Y}{n}}$$

$Z \sim N$, $Y \sim X$, n = degrees of freedom

Formula

Chi-Square Distribution

$$X \sim X$$

$$f(x) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{n}}}{\Gamma(\frac{n}{2})}, x > 0, n = \text{positive integer and degrees of freedom}$$

Formula

F Distribution

$$X \sim F_{n,d}$$

n = degrees of freedom for the numerator

d = degrees of freedom for the denominator

$$f(x) = \frac{\Gamma(\frac{u+v}{2})}{\Gamma(\frac{u}{2}) \Gamma(\frac{v}{2})} \frac{u^{\frac{u}{2}-1} x^{\frac{u}{2}-1}}{v^{\frac{u}{2}-1} x^{\frac{u}{2}-1}} \frac{v^{\frac{v}{2}-1} x^{\frac{v}{2}-1}}{x^{\frac{u}{2}-1} x^{\frac{v}{2}-1}}$$

$$X = \frac{Y_u}{W_v}, Y, W \text{ are chi-square}$$

Notes for the TI-83, 83+, 84 Calculator



Notes and tips for using TI-83, TI-83+, and TI-84 calculators for statistics applications.

Quick Tips

Legend

- 

represents a button press

-  represents yellow command or green letter behind a key
-  represents items on the screen

To adjust the contrast

Press



, then hold



to increase the contrast or



to decrease the contrast.

To capitalize letters and words

Press



to get one capital letter, or press



, then

ALPHA

to set all button presses to capital letters. You can return to the top-level button values by pressing

ALPHA

again.

To correct a mistake

If you hit a wrong button, just hit

CLEAR

and start again.

To write in scientific notation

Numbers in scientific notation are expressed on the TI-83, 83+, and 84 using E notation, such that...

- $4.321 \text{ E } 4 = 4.321 \times 10^4$
- $4.321 \text{ E } -4 = 4.321 \times 10^{-4}$

To transfer programs or equations from one calculator to another:

Both calculators: Insert your respective end of the link cable and press

2nd

, then **[LINK]**.

Calculator receiving information:


Use the arrows to navigate to and select **<RECEIVE>**

Press

ENTER

Calculator sending information:

Press appropriate number or letter.

Use up and down arrows to access the appropriate item.
Press  to select item to transfer.

Press right arrow to navigate to and select **<TRANSMIT>**.
Press .

Note:ERROR 35 LINK generally means that the cables have not been inserted far enough.

Both calculators: Insert your respective end of the link cable cable Both calculators: press



, then **[QUIT]** To exit when done.

Manipulating One-Variable Statistics

Note:These directions are for entering data with the built-in statistical program.

Data	Frequency
-2	10

Data	Frequency
-1	3
0	4
1	5
3	8

Sample Data We are manipulating 1-variable statistics.

To begin:

Turn on the calculator.

ON

Access statistics mode.

STAT

Select **<4:ClrList>** to clear data from lists, if desired.

4

,

ENTER

Enter list **[L1]** to be cleared.

2nd

, [L1] ,

ENTER

Display last instruction.

2nd

, [ENTRY]

Continue clearing remaining lists in the same fashion, if desired.

◀

,

2nd

, [L2],

ENTER

Access statistics mode.

STAT

Select<1:Edit . . .>

ENTER

Enter [L1]. (You may [L1])
data. Data need to
values go arrow over
into to

- Type in a data value and enter it. (For negative numbers, use the negate (-) key at the bottom of the keypad)

(-)

,

—

In [L2], enter the frequencies for each data value in

[L1].

- Continue in the same manner until all data values are entered.

- Type in a frequency and enter it. (If a data value appears only once, the frequency is "1")

- Continue in the same manner until all data values are entered.

Access statistics mode.

Navigate to <CALC>
Access <1:1-var Stats>

Indicate that the data is in [L1]...

2nd
, [L1] ,

...and indicate that the frequencies are in **[L2]**.

2nd

, **[L2]** ,

ENTER

The statistics should be displayed. You may arrow down to get remaining statistics. Repeat as necessary.

Drawing Histograms

Note: We will assume that the data is already entered

We will construct 2 histograms with the built-in STATPLOT application. The first way will use the default ZOOM. The second way will involve customizing a new graph.

Access graphing mode.

2nd

, **[STAT PLOT]**

Select **<1:plot 1>** To access plotting - first graph.

ENTER

Use the arrows navigate go to **<ON>** to turn on Plot 1.

<ON> .



ENTER

Use the arrows to go to the histogram picture and select the histogram.

ENTER

Use the arrows to navigate to **<Xlist>**
If "L1" is not selected, select it.

2nd

, [L1],

ENTER

Use the arrows to navigate to **<Freq>**.
Assign the frequencies to **[L2]**.

2nd

, [L2],

ENTER

Go back to access other graphs.

2nd

, [STAT PLOT]

Use the arrows to turn off the remaining plots.

Be sure to deselect or clear all equations before graphing.

To deselect equations:

Access the list of equations.

Y=

Select each equal sign (=).



Continue, until all equations are deselected.

To clear equations:

Access the list of equations.



Use the arrow keys to navigate to the right of each equal sign (=) and clear them.



Repeat until all equations are deleted.

To draw default histogram:

Access the ZOOM menu.



Select **<9:ZoomStat>**




The histogram will show with a window automatically set.

To draw custom histogram:

Access  to set the graph parameters.

- $X_{\min} = -2.5$
- $X_{\max} = 3.5$
- $X_{\text{scl}} = 1$ (width of bars)
- $Y_{\min} = 0$
- $Y_{\max} = 10$
- $Y_{\text{scl}} = 1$ (spacing of tick marks on y-axis)
- $X_{\text{res}} = 1$

Access  to see the histogram.

To draw box plots:

Access graphing mode.



, 

Select  to access the first graph.

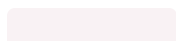


Use the arrows to select  and turn on Plot 1.



Use the arrows to select the box plot picture and enable it.





Use the arrows to navigate to **<Xlist>**
If "L1" is not selected, select it.

2nd
, [L1] ,
ENTER

Use the arrows to navigate to **<Freq>**.
Indicate that the frequencies are in **[L2]**.

2nd
, [L2] ,
ENTER

Go back to access other graphs.

2nd
, [STAT PLOT]

Be sure to deselect or clear all equations before graphing using the method mentioned above.
View the box plot.

GRAPH
, [STAT PLOT]

Linear Regression

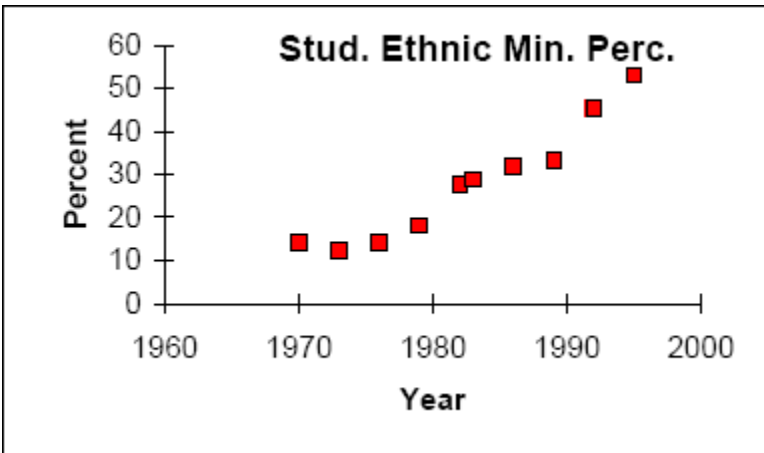
Sample Data

The following data is real. The percent of declared ethnic minority students at De Anza College for selected years from 1970 - 1995 was:

Year	Student Ethnic Minority Percentage
1970	14.13
1973	12.27
1976	14.08
1979	18.16
1982	27.64
1983	28.72
1986	31.86
1989	33.14
1992	45.37
1995	53.1

The independent variable is "Year," while the dependent variable is "Student Ethnic Minority Percent."

Student Ethnic Minority Percentage



By hand, verify the scatterplot above.

Note: The TI-83 has a built-in linear regression feature, which allows the data to be edited. The x-values will be in

[L1]

; the y-values in

[L2]

.

To enter data and do linear regression:

ON Turns calculator on

ON

Before accessing this program, be sure to turn off all plots.

- Access graphing mode

mode.

2nd

, **[STAT PLOT]**

- Turn off all plots.

4

,

ENTER

Round to 3 decimal places. To do so:

- Access the mode menu.

MODE

, **[STAT PLOT]**

- Navigate to **<Float>** and then to the right to **<3>**.

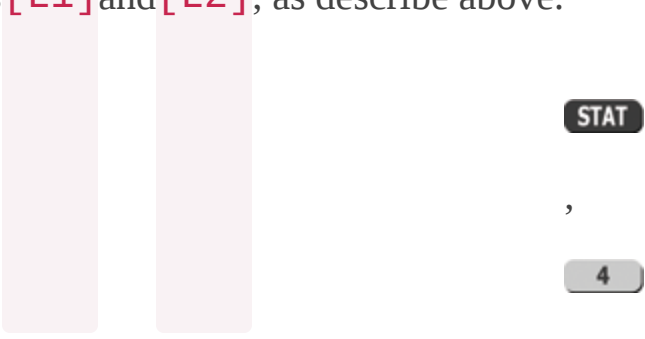
▼

▶

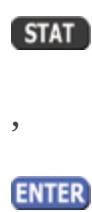
- All numbers will be rounded to 3 decimal places until changed.

ENTER

Enter statistics mode and clear lists [L1] and [L2], as describe above.



Enter editing mode to insert values for x and y.



Enter each value. Press



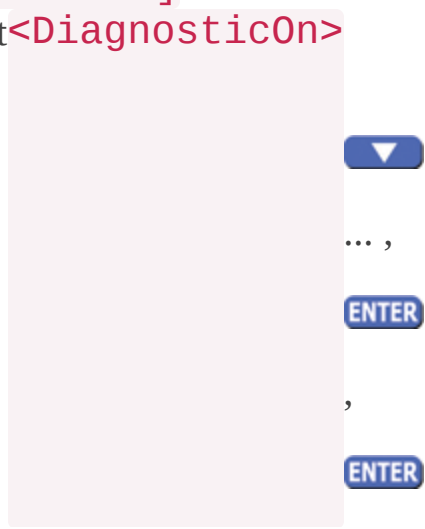
To display the correlation coefficient:

Access the catalog.



, [CATALOG]

Arrow down and select<DiagnosticOn>



r and r^2 will be displayed during regression calculations.
Access linear regression.

STAT



Select the form of $y = a + bx$

8

,

ENTER

The display will show:

LinReg

- $y = a + bx$
- $a = -3176.909$
- $b = 1.617$
- $r^2 = 0.924$
- $r = 0.961$

This means the Line of Best Fit (Least Squares Line) is:

- $y = -3176.909 + 1.617x$
- $\text{Percent} = -3176.909 + 1.617(\text{year } \#)$

The correlation coefficient $r = 0.961$

To see the scatter plot:

Access graphing mode.

2nd

, [STAT PLOT]

Select **<1:plot 1>** To access plotting - first graph.

ENTER

Navigate and select **<ON>** to turn on Plot 1.

<ON>

ENTER

Navigate to the first picture.
Select the scatter plot.

ENTER

Navigate to **<Xlist>**

If **[L1]** is not selected, press **2nd**, **[L1]** to select it.

2nd

Confirm that the data values are in **[L1]**.

<ON>

ENTER

Navigate to **<Ylist>**

Select that the frequencies are in **[L2]**.

2nd

, **[L2]** ,

ENTER

Go back to access other graphs.

2nd

[STAT PLOT]


, [STAT PLOT]

Use the arrows to turn off the remaining plots.

Access  to set the graph parameters.

- $X_{\min} = 1970$
- $X_{\max} = 2000$
- $X_{\text{scl}} = 10$ (spacing of tick marks on x-axis)
- $Y_{\min} = -0.05$
- $Y_{\max} = 60$
- $Y_{\text{scl}} = 10$ (spacing of tick marks on y-axis)
- $X_{\text{res}} = 1$

Be sure to deselect or clear all equations before graphing, using the instructions above.

Press  to see the scatter plot.

To see the regression graph:

Access the equation menu. The regression equation will be put into Y1.

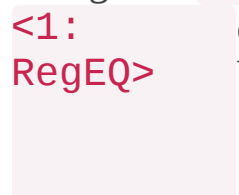


Access the vars menu and navigate to 





Navigate to .

 contains the regression equation which will be entered in Y1.



Press **GRAPH**. The regression line will be superimposed over scatter plot.

To see the residuals and use them to calculate the critical point for an outlier:

Access the list. RESID will be an item on the menu.
Navigate to it.

2nd
, **[LIST]**,
<RESID>

Confirm twice to view the list of residuals. Use the arrows to select them.

ENTER

,

ENTER

The critical point for an outlier is: $1.9V \frac{SSE}{n-2}$ where:

- n = number of pairs of data
- SSE = sum of the squared errors
- residual²

Store the residuals in **[L3]**.

STO►

,

2nd

, **[L3]** ,

ENTER

Calculate the $\frac{(\text{residual})^2}{n-2}$. Note that $n - 2 = 8$

2nd

, [L3],

x²

,

÷

,

8

Store this value in [L4].

STO▶

,

2nd

, [L4],

ENTER

Calculate the critical value using the equation above.

1

,

.

,

9

,



,



, [V] ,



, [LIST]



,



,



,



, [L4] ,



,



,



Verify that the calculator displays: 7.642669563. This is the critical value.

Compare the absolute value of each residual value in [L3] to 7.64 . If the absolute value is greater than 7.64, then the (x, y) corresponding point is an outlier. In this case, none of the points is an outlier.

To obtain estimates of y for various x-values:

There are various ways to determine estimates for "y". One way is to substitute values for "x" in the equation. Another way is to use the

TRACE

on the graph of the regression line.

TI-83, 83+, 84 instructions for distributions and tests

Distributions

Access **DISTR** (for "Distributions").

For technical assistance, visit the Texas Instruments website at <http://www.ti.com> and enter your calculator model into the "search" box.

Binomial Distribution

- **binompdf(n, p, x)** corresponds to $P(X = x)$
- **binomcdf(n, p, x)** corresponds to $P(X \leq x)$
- To see a list of all probabilities for $x: 0, 1, \dots, n$, leave off the "x" parameter.

Poisson Distribution

- **poissonpdf(λ , x)** corresponds to $P(X = x)$
- **poissoncdf(λ , x)** corresponds to $P(X \leq x)$

Continuous Distributions (general)

- $-\infty$ uses the value -1EE99 for left bound
- ∞ uses the value 1EE99 for right bound

Normal Distribution

- `normalpdf(x, μ , σ)` yields a probability density function value (only useful to plot the normal curve, in which case " x " is the variable)
- `normalcdf(left bound, right bound, μ , σ)` corresponds to $P(\text{left bound} < X < \text{right bound})$
- `normalcdf(left bound, right bound)` corresponds to $P(\text{left bound} < Z < \text{right bound})$ - standard normal
- `invNorm(p, μ , σ)` yields the critical value, k : $P(X < k) = p$
- `invNorm(p)` yields the critical value, k : $P(Z < k) = p$ for the standard normal

Student-t Distribution

- `tpdf(x, df)` yields the probability density function value (only useful to plot the student-t curve, in which case " x " is the variable)
- `tcdf(left bound, right bound, df)` corresponds to $P(\text{left bound} < t < \text{right bound})$

Chi-square Distribution

- `X2pdf(x, df)` yields the probability density function value (only useful to plot the χ^2 curve, in which case " x " is the variable)
- `X2cdf(left bound, right bound, df)` corresponds to $P(\text{left bound} < X^2 < \text{right bound})$

F Distribution

- `Fpdf(x, dfnum, dfdenom)` yields the probability density function value (only useful to plot the F curve, in which case " x " is the variable)
- `Fcdf(left bound, right bound, dfnum, dfdenom)` corresponds to $P(\text{left bound} < F < \text{right bound})$

Tests and Confidence Intervals

Access **STAT** and **TESTS**.

For the Confidence Intervals and Hypothesis Tests, you may enter the data into the appropriate lists and press **DATA** to have the calculator find the sample means and standard deviations. Or, you may enter the sample means and sample standard deviations directly by pressing **STAT** once in the appropriate tests.

Confidence Intervals

- **ZInterval** is the confidence interval for mean when σ is known
- **TInterval** is the confidence interval for mean when σ is unknown; s estimates σ .
- **1-PropZInt** is the confidence interval for proportion

Note: The confidence levels should be given as percents (ex. enter "**95**" or "**.95**" for a 95% confidence level).

Hypothesis Tests

- **Z-Test** is the hypothesis test for single mean when σ is known
- **T-Test** is the hypothesis test for single mean when σ is unknown; s estimates σ .
- **2-SampZTest** is the hypothesis test for 2 independent means when both σ 's are known
- **2-SampTTest** is the hypothesis test for 2 independent means when both σ 's are unknown
- **1-PropZTest** is the hypothesis test for single proportion.
- **2-PropZTest** is the hypothesis test for 2 proportions.
- **χ^2 -Test** is the hypothesis test for independence.
- **χ^2 GOF-Test** is the hypothesis test for goodness-of-fit (TI-84+ only).
- **LinRegTTEST** is the hypothesis test for Linear Regression (TI-84+ only).

Note: Input the null hypothesis value in the row below "**Inpt.**" For a test of a single mean, " μ_0 " represents the null hypothesis. For a test of a single proportion, " p_0 " represents the null hypothesis. Enter the alternate hypothesis on the bottom row.

Tables

Note: When you are finished with the table link, use the back button on your browser to return here.

Tables (NIST/SEMATECH e-Handbook of Statistical Methods, <http://www.itl.nist.gov/div898/handbook/>, January 3, 2009)

- [Student-t table](#)
- [Normal table](#)
- [Chi-Square table](#)
- [F-table](#)
- All four tables can be accessed by going to <http://www.itl.nist.gov/div898/handbook/eda/section3/eda367.htm>

95% Critical Values of the Sample Correlation Coefficient Table

- [95% Critical Values of the Sample Correlation Coefficient](#)

Note: The url for this table is <http://cnx.org/content/m17098/latest/>